Amino acid substitutions in protein binding: A study for peptides and antibodies

DISSERTATION

zur Erlangung des akademischen Grades
doctor rerum naturalium
(Dr. rer. nat.)
im Fach Biologie

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät I
Humboldt-Universität zu Berlin

von
Herrn Dipl.-Math. Armin A. Weiser
geboren am 11.12.1969 in Bad Hersfeld

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Christoph Markschies

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I:
Prof. Dr. Lutz-Helmut Schön

Gutachter:

1. Dr. Michal Or-Guil
2. Prof. Cornelius Frömmel
3. Prof. Sepp Hochreiter

Tag der mündlichen Prüfung: 29. Juli 2009

Abstract

A central task of the evolutionary process is the alteration of amino acid sequences, such as the substitution of one amino acid by another. Not only do these amino acid changes occur gradually over large time scales and result in the variety of life surrounding us, but they also happen daily within an organism. Such alterations take place rapidly for the purposes of defense, which in higher vertebrates, is managed by the humoral immune system. For an effective immune response, antibodies are subjected to a micro-evolutionary process that includes multiple rounds of diversification by somatic hypermutation resulting in increased binding affinity to a particular pathogen. In contrast to recent advances in revealing the molecular mechanism of somatic hypermutation, the fundamentals of the selection process are still poorly understood. The goal of this work was to provide insights into the micro-evolution of antibodies during the immune response, including the relationship between amino acid substitutions and binding affinity changes.

A preliminary step in this work was to understand how to reliably quantify the outcome of the SPOT synthesis (positionally addressable synthesis of peptides on cellulose membranes). Using the monoclonal antibody CB4-1 as a model system, dissociation constants of antibody-peptide complexes were systematically compared to the signal intensities obtained by the SPOT synthesis. By analyzing a set of peptides possessing different affinities to CB4-1, a model for the relationship between the dissociation constant and signal intensity based on the mass action law could be developed. The accuracy of the SPOT screening method was determined and the experimental requirements needed to obtain reliable binding affinity class predictions were discussed. It could be shown that the SPOT synthesis is an accurate method for assigning measured signal intensities to three different binding affinity classes.

A substitution matrix based on data produced with these binding experiments was constructed and named AFFI. AFFI is the first substitution matrix that is based solely on binding affinity and AFFI has the highest prediction power for binding affinity data compared to other commonly used substitution matrices. Application of a clustering algorithm based on simulated annealing permitted the use of AFFI to construct optimal clusters of amino acids. As revealed by a theoretical approach, an optimal, reduced set of amino acids could be provided for the purposes of epitope searching.

Furthermore, AFFI was used to investigate the search strategy of sequence space applied by the immune system. The structure of an antibody exhibits significant differences in its underlying sequences, indicative

of a coarse coverage of sequence space *a priori* (before any contact to a pathogen). For the *a posteriori* process of somatic hypermutation and selection, a novel approach to identify mutations relevant to affinity maturation is presented. This uses a statistical assessment of the frequency distribution of mutations observed in a large dataset of $VH_{186.2}$ sequences originating from previous investigations of the primary response to the hapten 4-hydroxy-3-nitrophenylacetyl (NP) in C57BL/6 mice. For the statistical investigation presented here, it is necessary that all sequences are independent of each other, meaning they must stem from different clones. Due to the lack of a method for this purpose, a novel approach was developed here that should be useful for related applications as well. Various mutations that are favored by the selection mechanism could be identified in the clonally unrelated $VH_{186.2}$ sequences, which occur significantly more often than predicted by intrinsic mutability. The analysis revealed that the spectrum of mutations favored by the selection process is much broader than previously thought. The fact that particular silent mutations are strongly favored indicates either that intrinsic mutability has been grossly underestimated, or that selection acts not only on antibody affinity but also on their expression rates, making affinity maturation and "expression rate maturation" concurrent processes. The implications of these results are discussed in the context of current models of affinity maturation.

Keywords:
peptide array, antibody, substitution matrix, affinity maturation

Zusammenfassung

Die strukturelle und somit funktionelle Modifizierung von Proteinsequenzen unter anderem durch den Austausch von Aminosäuren ist ein zentraler Aspekt in evolutionären Prozessen. Solche Prozesse ereignen sich jedoch nicht nur innerhalb großer Zeiträume und resultieren in der Vielfalt des Lebens, das uns umgibt, sondern sind auch täglich innerhalb von Organismen beobachtbar. Diese mikroevolutionären Prozesse finden mit hoher Geschwindigkeit statt. Sie bilden eine Grundlage zur Immunabwehr höherer Wirbeltiere und werden durch das humorale Immunsystem organisiert. So werden im Zuge einer Immunantwort Antikörper wiederholt der Diversifizierung durch somatische Hypermutation unterworfen. Dieses führt zu einer verbesserten Bindungsaffinität des Antikörpers zum Pathogen und somit zu dessen effektiveren Beseitigung. In der Aufklärung der molekularen Mechanismen der somatischen Hypermutation wurden große Fortschritte gemacht. Die Grundlagen des Selektionsprozesses hingegen werden noch unzureichend verstanden. Ziele dieser Arbeit waren, neue Kenntnisse über die Mikroevolution von Antikörpern während der Immunantwort zu gewinnen und die Beziehung zwischen Aminosäureaustauschen und Affinitätsänderungen zu verstehen.

Ein vorbereitender Schritt dieser Arbeit war es, die Ergebnisse der SPOT Synthese (Synthese von Peptiden auf Zellulosemembranen in definierten Positionen) zuverlässig quantifizieren zu können. Hierzu wurde der monoklonale Antikörper CB4-1 als Modellsystem verwendet. Bekannte Dissoziationskonstanten von Antikörper-Peptid Komplexen wurden systematisch mit den Signalintensitäten der SPOT Synthese verglichen. Durch die Analyse von Peptiden mit unterschiedlichen Affinitäten zu CB4-1 konnte ein auf dem Massenwirkungsgesetz basierendes Modell entwickelt werden. Das Modell stellt den Zusammenhang zwischen Signalintensitäten und Dissoziationskonstanten dar. Weiterhin konnten die Genauigkeit der durch die SPOT Synthese bestimmten Bindungsaffinitäten bestimmt werden und experimentelle Anforderungen definiert werden, um zuverlässig messbare Bindungsaffinitätsklassen abgrenzen zu können. Es konnte gezeigt werden, dass die SPOT Synthese eine präzise Methode ist, um Signalintensitäten drei verschiedenen Bindungsaffinitätsklassen zuzuordnen.

Antikörper-Peptid Bindungsdaten, die aus SPOT Synthese Experimenten generiert wurden, bildeten die Grundlage zur Konstruktion der Substitutionsmatrix AFFI - der ersten Substitutionsmatrix, die ausschließlich auf Bindungsaffinitätsdaten beruht. Im Vergleich zu anderen weit verbreiteten

Substitutionsmatrizen kann AFFI Bindungsaffinitätsänderungen am verlässlichsten vorhersagen. Mit Hilfe von AFFI und einem Simulated Annealing Algorithmus wurden Aminosäuren optimal gruppiert. Ein reduzierter Aminosäuresatz wurde gewonnen. Durch einen theoretischen Ansatz konnte gezeigt werden, dass der reduzierte Aminosäuresatz eine optimale Basis für die Epitopsuche darstellt.

Darüber hinaus konnten mittels AFFI die Strategien des Immunsystems zur Abdeckung des Sequenzraumes untersucht werden. Verschiedene Antikörper zeigen bedeutende Unterschiede in ihren zugrunde liegenden Sequenzen, die eine grobe Abdeckung des Sequenzraums *a priori* (vor dem ersten Kontakt zu einem Pathogen) ermöglichen. Für den *a posteriori* Prozess der somatischen Hypermutation und Selektion wird ein neuer Ansatz präsentiert, um für die Affinitätsreifung relevante Mutationen zu identifizieren. Dafür wurde die primäre Immunantwort von C57BL/6 Mäusen auf das Hapten 4-hydroxy-3-nitrophenylacetyl (NP) als Modellsystem verwendet. Aus verschiedenen Publikationen wurde ein großer Datensatz der dazugehörigen $VH_{186.2}$ Sequenzen extrahiert und die Häufigkeitsverteilung von Mutationen statistisch ausgewertet. Für die hier präsentierte statistische Untersuchung ist es notwendig, dass alle Sequenzen voneinander unabhängig sind, d.h. dass sie von verschiedenen Klonen stammen müssen. Um die klonale Unabhängigkeit von Sequenzen ermitteln zu können, wurde eine neue Methode entwickelt, die auch für ähnliche Fragestellungen äußerst hilfreich sein kann. In den von unterschiedlichen Klonen stammenden $VH_{186.2}$ Sequenzen konnten mehrere Mutationen identifiziert werden, die häufiger vorkommen als durch intrinsische Mutabilität erwartet. Sie wurden durch den Selektionsmechanismus bevorzugt. Die Analyse zeigte, dass das Spektrum der selektierten Mutationen viel umfangreicher ist als bisher angenommen wurde. Die Tatsache, dass auch einige stille Mutationen stark bevorzugt werden, deutet darauf hin, dass entweder die intrinsische Mutabilität stark unterschätzt wurde oder, dass Selektion nicht nur auf Affinitätsreifung von Antikörpern basiert sondern auch auf ihrer Expressionsrate. Letzteres würde bedeuten, dass Affinitätsreifung und "Expressionsratenreifung" simultan ablaufende Prozesse sind. Die Implikationen dieser Ergebnisse werden im Zusammenhang mit aktuellen Modellen der Affinitätsreifung diskutiert.

Schlagwörter:
Peptidarray, Antikörper, Substitutionsmatrix, Affinitätsreifung

Contents

List of Figures

List of Tables

Chapter 1

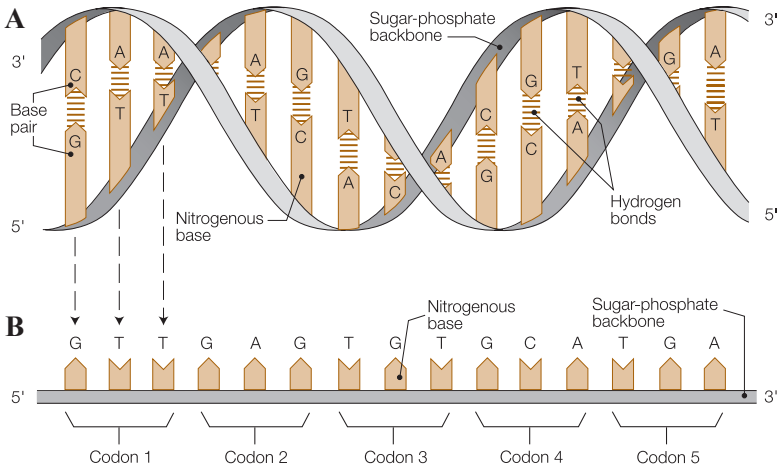
Introduction

1.1 DNA

Each cell contains the entire genetic information necessary for the development and functioning of a living organism. This information is stored in the deoxyribonucleic acid (DNA). A DNA is a polymeric molecule composed of a sequence of monomeric subunits called nucleotides. The nucleotides consist of the sugar 2-deoxyribose, a phosphate group and one out of four nitrogenous bases: adenosine (A) and guanosine (G) (purine derivatives), and thymidine (T) and cytidine (C) (pyrimidine derivatives). Nucleotides are linked together by phosphodiester bonds between the 5' and the 3' carbon atoms of adjacent nucleotides. Conventionally, a DNA molecule is defined to start at its 5' end and to finish at its 3' end (Figure 1.1).

DNA consists of two antiparallel strands. They form a double helix, which is stabilized by characteristic hydrogen bonds between the bases of the two strands. The interaction of bases is very specific such that the two DNA strands are complementary. Adenine only forms hydrogen bonds with thymine, whereas cytosine only interacts with guanine (Figure 1.1).

For eukaryotic organisms the DNA resides in the nucleus of a cell within the chromosomes. These chromosomes are duplicated before cells divide and the DNA is replicated by uncoiling the double helix. The separated complementary strands serve as a template for DNA polymerases, which synthesize new DNA strands. The term, gene, is used for a region of the DNA that encodes for a regulatory function of gene expression, for proteins or for RNA (ribonucleic acid) molecules.



Source: *Molecular Biology of the Cell*, Garland Publishing Inc., New York 1994

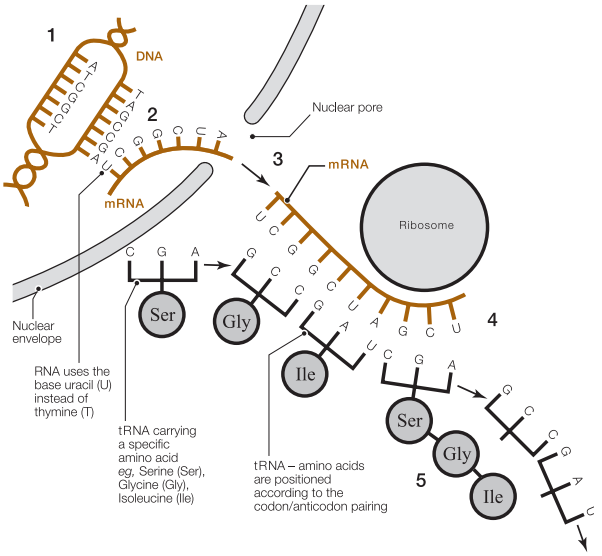
Figure 1.1: DNA. (A) DNA double helix section; (B) Schematic representation of a single (coding) strand of DNA.

1.1.1 Transcription

The transcription of DNA is the first step in the expression of a protein coding gene. In this process a DNA sequence is transcribed into a single stranded messenger RNA (mRNA). Compared to DNA, the mRNA backbone is built out of ribose instead of deoxyribose, and thymine (T) is replaced by uracil (U).

Similar to DNA replication, the RNA is synthesized from the 5' to the 3' end. Therefore, one DNA strand - the template strand - is transcribed into mRNA, which is catalyzed by RNA polymerase II. The other DNA strand that is not transcribed is called the coding strand, because it contains the same coding sequence as the RNA transcript (with the exception of thymine instead of uracil nucleotides).

The resulting mRNA is initially called pre-mRNA, because it is further processed in the nucleus. Pre-mRNA consists of so-called introns and exons. The introns are DNA regions within a gene that are usually not translated into a protein and therefore are removed from the pre-mRNA by a process called splicing. After splicing the mRNA consists only of exons. Finally, the mature mRNA migrates out of the nucleus destined for translation (Figure 1.2).



Source: *Human Biology and Health Studies*, Thomas Nelson, Walton-on-Thames, 1996

Figure 1.2: Protein synthesis. 1. The DNA double helix unwinds and exposes a nucleotide sequence. 2. One of the strands is transcribed into mRNA which travels out of the nucleus of the cell. 3. The mRNA couples with the ribosome. 4. Translation is assisted by tRNA molecules carrying amino acids to the ribosome, which adds them to the growing polypeptide chain. 5. As the polypeptide chain grows, it folds into a protein.

1.1.2 Translation

The translation of mRNAs is processed by ribosomes, complexes composed of ribosomal RNA (rRNA) and ribosomal proteins, where mRNA, synthesized during transcription, is translated into the corresponding polypeptide chain. The ribosome reads triplets of nucleotides encoded in the mRNA called codons, each mapping to a specific amino acid. This genetic code is given in Table 1.1. There are $4^3 = 64$ possible nucleotide triplets. Since 61 codons specify only 20 amino acids, the genetic code is redundant. The remaining three triplets are STOP codons. A position of a codon is said to be a n-fold degenerate site if n nucleotides at this position specify the same amino acid. Transfer RNAs (tRNA) are the interface between a codon and an amino acid. The concentration of tRNA in a cell is given by the codon usage (Table 1.2) and differs from species to species.

The start codon (AUG) initializes the translation of mRNA. While mRNA is shifted through the ribosome, tRNA-attached amino acids are released and catalytically added to the growing peptide chain. When the ribosome recognizes one of the three stop-codons, there is no associated tRNA and consequently the translation is terminated. Finally, the mRNA and the polypeptide are released from the ribosome (Figure 1.2).

1.1.3 Polypeptides

Polypeptides and proteins are involved in every process within a cell and are therefore essential in every living organism. The building blocks of polypeptides are amino acids, which consist of a central alpha Carbon (C_α) to which an amino group (NH_2), a carboxyl group (COOH) and a variable side chain (R_i) are bound. There are 20 different amino acids, each made up of different physical and chemical properties (Figure 1.5) depending on the constitution of their side chain, reaching from non-polar hydrophobic amino acids like alanine (R_i=-CH_3) to polar or even charged amino acids like aspartic acid (R_i=-CH_2-COOH). A polypeptide is a single linear chain of amino acids. Two amino acids are connected to form a dipeptide, up to 50 connected amino acids are called peptides and even more form a protein, respectively.

Under dehydration the amino group of one amino acid and the carboxyl group of another amino acid are joined together via a peptide bond (Figure 1.3). The linear amino acid sequence, also referred to as the primary structure of the protein is conventionally written from the amino-terminus (N-terminus) to the carboxyl-terminus (C-terminus). To fulfill their function(s) proteins form secondary structures, stabilized by hydrogen bonds, and fold to three-dimensional structures (tertiary structures).

First Base	Second Base: U	Second Base: C	Second Base: A	Second Base: G	Third Base
U	Phenylalanine	Serine	Tyrosine	Cysteine	**U**
					C
	Leucine		STOP	STOP	**A**
				Tryptophan	**G**
C	Leucine	Proline	Histidine	Arginine	**U**
					C
			Glutamine		**A**
					G
A	Isoleucine	Threonine	Asparagine	Serine	**U**
					C
			Lysine	Arginine	**A**
	Methionine - START				**G**
G	Valine	Alanine	Aspartic Acid	Glycine	**U**
					C
			Glutamic Acid		**A**
					G

Table 1.1: The genetic code.

First Base	Second Base: U	Second Base: C	Second Base: A	Second Base: G	Third Base
U	17.2	16.2	12.2	11.4	**U**
	21.8	18.1	16.1	12.3	**C**
	6.7	11.8	1.0	1.6	**A**
	13.4	4.2	0.8	12.5	**G**
C	13.4	18.4	10.6	4.7	**U**
	20.2	18.2	15.3	9.4	**C**
	8.1	17.3	12.0	6.6	**A**
	39.5	6.2	34.1	10.2	**G**
A	15.4	13.7	15.6	12.7	**U**
	22.5	19.0	20.3	19.7	**C**
	7.4	16.0	21.9	12.1	**A**
	22.8	5.6	33.6	12.2	**G**
G	10.7	20.0	21.0	11.4	**U**
	15.4	26.0	26.0	21.2	**C**
	7.4	15.8	27.0	16.8	**A**
	28.4	6.4	29.4	15.2	**G**

Table 1.2: The codon usage of the house mouse (*mus musculus*)[a].

[a]The data was retrieved from the Codon Usage Database http://www.kazusa.or.jp/codon. The frequency is given per thousand.

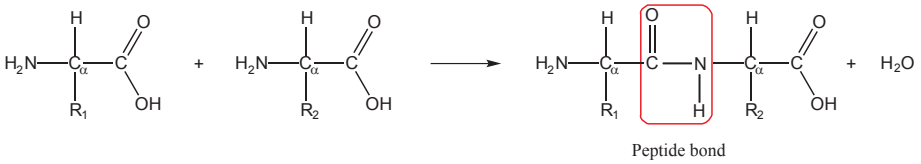


Figure 1.3: The peptide bond. Two amino acids are linked by dehydration synthesis to form a dipeptide.

1.1.4 Point mutations

A point mutation in DNA is a change of one single nucleotide. This may be an insertion of a new, a substitution of an existing nucleotide by another, or a deletion of one. Mutations can be due to errors during DNA replication or due to mutagens. These are usually chemical agents or ionizing radiation like UV rays or X-rays.

The most frequently occurring type of nucleotide substitutions are transitions where a purine nucleotide is replaced by another purine nucleotide (A,G), or a pyrimidine nucleotide is replaced by another pyrimidine nucleotide (C,T). Substitutions of a purine by a pyrimidine nucleotide or vice versa are called transversions (Figure 1.4).

Nucleotide substitutions in protein coding sequences are classified according to their effect on the amino acid sequence. While silent or synonymous substitutions do not cause a change of the specified amino acid sequence, replacement or nonsynonymous substitutions alter it. Consequently, a nonsynonymous substitution results in the replacement of one amino acid by another. A special type of nonsynonymous substitution is the nonsense mutation. These nonsense mutations change an amino acid codon into a stop codon leading to the truncation of the protein sequence.

1.2 Amino acid sequence space

1.2.1 Evolutionary theories

According to Darwin's theory of *evolution by natural selection*, mutations in individuals are positively selected and fixed in a population if they improve the individual's fitness, their ability to survive and to reproduce.

Kimura estimated that the rate of amino acid replacements was too high to be explained by theories of natural selection [65]. This caused him to postulate the *neutral theory of evolution* - the random fixation of neutral mutations that do not affect the fitness of an individual. Later, DNA sequencing technology revealed that the rate of synonymous substitutions is

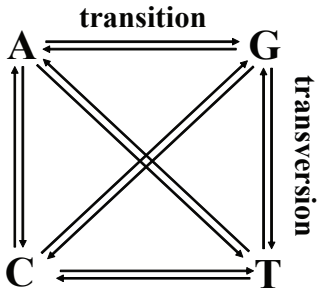


Figure 1.4: Scheme of the transition/transversion bias. A transition conserves the structural features of the nucleotide. This means that a purine nucleotide remains a purine nucleotide (A, G) and a pyrimidine nucleotide mutates into a pyrimidine nucleotide (T, C). In case of a transversion, a purine nucleotide is substituted by a pyrimidine nucleotide and vice versa.

much larger than the rate of nonsynonymous substitutions and that Kimura even underestimated the rate of evolution acting on the molecular level.

Today, modern evolutionary theories at the molecular level generally are consistent with both aspects of the evolutionary process, selection and neutrality.

If a mutation results in the exchange of one amino acid to a dissimilar one, in terms of physicochemical properties, the protein may loose the ability to fold and to take a similar or modified function. Consequently, the fitness of the organism is reduced and the organism has a high chance to be removed from the population. Such a mutation is called a deleterious mutation and the type of selection is called negative or purifying selection. In contrast, synonymous substitutions are expected to be selectively neutral and exchanges of amino acids with similar chemical and physical properties are expected to be either neutral or advantageous. Dayhoff revealed the concrete patterns of amino acid replacements and tabulated frequencies and types of amino acid replacements in the 1970ies [27]. She found that the acceptance of an amino acid replacement does not primarily depend on the number of nucleotide substitutions that are required to interchange an amino acid into another. Preferentially, physicochemically similar amino acids are exchanged. Her results revealed that natural selection operates on amino acid replacements.

1.2.2 Amino acid properties

The 20 types of amino acids introduce not only diversity and complexity into proteins, but also some specific propensities. Comparison of physicochemical

properties of amino acids enables the classification into different physicochemically similar groups [63], as some amino acids are similar in physicochemical properties (Figure 1.5) and exchanges of these amino acids can be tolerated in many regions of a sequence [126]. However, such physicochemical classification systems are very poor at predicting possible replacements of a sequence position by other amino acids. In fact, the exchangeability of a specific sequence position is ruled by the sequence position requirements formed by the sum of all molecular interactions taking place at that position. This includes the interactions necessary for structural stability as well as those important for ligand binding. Therefore, substitution probabilities might differ depending on the matter of the observation. The exchangeability within the framework of folded proteins might be different on the surface or binding region of proteins like the complementarity-determining regions (CDR) of antibodies (section 1.3.1). Additionally, the underlying gene sequences of proteins affect the exchangeability of amino acids within short time scales (i.e., few generations) e.g., during affinity maturation of antibodies [119].

An example for the tightrope walk of mutations is the sickle-cell disease. The gene defect is a known mutation of a single nucleotide (adenine to thymine) of the β-globin gene, which results in glutamate to valine substitution at position six of the β-subunit of hemoglobin. The affected red blood cells assume a sickle shape, when the oxygen partial pressure decreases. They get easily stuck in the capillaries and are often lysed. On the other hand, affected persons are resistant to malaria, hence this mutation is widespread in malarial territories. Another example where single mutations have beneficial

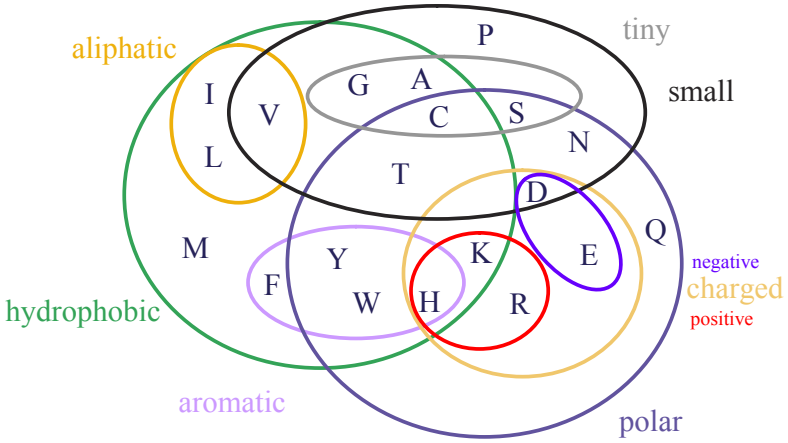


Figure 1.5: Amino acid properties. Grouping of amino acids according to their properties adapted from Livingstone and Barton [83].

effects are the so-called key mutations, that arise during affinity maturation in the germinal center reaction (see below in section 1.3.3). A single mutation may result in an antibody that shows binding affinity enhancements to the pathogen in the range of one magnitude [3].

1.2.3 Amino acid substitution matrices

An amino acid substitution matrix describes the probability of an amino acid exchange between all possible amino acid pairs. Amino acid substitution matrices are usually described as log-odd matrices and do not only contain information about the exchange probability, but also the probability of occurrence of each mutation. One of the first amino acid substitution matrices, the PAM (Point Accepted Mutation) matrix was developed in the 1970s by Margaret O. Dayhoff [27]. In the last 40 years a vast amount of substitution matrices were developed [63], each based on different fundaments and/or samples. The most common matrix is the BLOSUM (BLOck SUbstitution Matrix) [52]. Dayhoff's methodology of comparing closely related species turned out to be less applicable for aligning evolutionarily divergent sequences. Sequence changes over long evolutionary time scales are not well approximated by compounding small changes that occur over short time scales. BLOSUM rectifies this problem; and consequently, the BLOSUM62 matrix is the most used in sequence alignment applications.

1.2.4 Reduction of the set of amino acids

A central task of protein sequence analysis is to uncover the exact nature of information encrypted in the primary structure. Reducing the set of amino acids in a way that only the most indispensable amino acids remain [80], reduces the complexity of protein sequences and thus provides a simplified opportunity to study structure-function relationships in proteins.

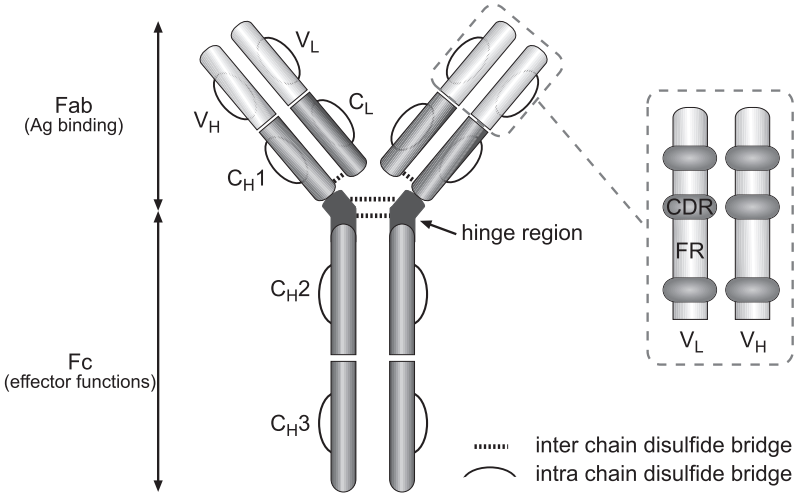
It has been discovered experimentally that designed proteins with fewer than 20 types of residues can have stable native structures [17, 26, 101, 106, 112]. From a physicist's perspective, this may imply that a 20 letter alphabet can be reduced into an N letter alphabet by partitioning the similar amino acids into N groups, and then N letters can be chosen as the representative residues of these N groups [17, 138]. Obviously, the simplest reduction is the so-called HP model [18, 74], where 20 types of amino acids are divided into two groups: H group and P group (H, hydrophobic residues; P, polar residues). Interestingly, such a type of simple two-letter HP model or the HP-like patterns could reproduce, to some extent, the kinetics and thermodynamics of protein folding and could be used to study the mechanism of

folding [26, 106]. Previously, a five-letter alphabet based on the statistical potential matrix by Miyazawa and Jernigan (MJ) - a pairwise interaction potential between amino acids [94] - was studied [17, 138]. Within this work five representative amino acids were given (Ile, Ala, Glu, Lys, Gly). This coincides with the experimental results obtained by Baker and co-workers [112]. A 57 residue SH3 domain with a β-barrel-like structure was studied, and 38 out of 40 targeted residues in the domain could be replaced with the same five types of amino acids. In further studies other simplified alphabets were described [21, 37, 38, 78, 80, 96, 128].

1.3 Affinity maturation in germinal centers

1.3.1 Antibodies - structure and function

Antibodies (Figure 1.6), also known as immunoglobulins (Ig), constitute a family of structurally related glycoproteins that participate in both the recognition and effector phases of the humoral immune response. In their membrane-bound form, antibodies are referred to as B cell antigen receptors (BCR). A typical human B cell will have $50,000$ to $100,000$ BCRs bound to its surface.



Source: Dissertation of Nicole Wittenbrink, HU Berlin, 2007

Figure 1.6: Schematic depiction of the Ig molecule exemplified for IgG. The antigen binding sites are formed by the variable light chain (V_L) and the variable heavy chain regions (V_H). C; constant region, CDR; complementarity determining region, Fab; fragment antigen binding, Fc; fragment crystallizable, FR; framework, H; heavy chain, L; light chain, V; variable region.

All antibodies share the same basic structure. The symmetric core structure of an antibody is a heterodimer, composed of two heavy and two light chains that are covalently connected via disulfide bonds (Figure 1.6). Each heavy and light chain consists of an amino-terminal variable (V) and a carboxy-terminal constant region (C). Due to differences in the constant region of the heavy chain, antibodies are subdivided into Ig classes (IgM, IgD, IgA, IgG and IgE) which permits different effector functions, such as complement activation and mediation of cell cytotoxicity. By contrast, the variable region designates the antigen specificity to the antibody and thereby accounts for recognition and binding of antigens, respectively. The antigen binding sites of an antibody are formed by the V region of one heavy (VH) and one light chain (VL). The variable regions feature three highly divergent stretches termed hypervariable segments that are positioned by conserved framework regions (FR) (Figure 1.6). In three-dimensional space, the three hypervariable segments of the (VH) chain and the three hypervariable segments of the (VL) chain are brought together to form the antigen binding site. Because the antigen binding site is complementary to the three-dimensional structure of bound antigen, the hypervariable segments are also referred to as complementarity-determining regions (CDR).

The unique part of the antigen recognized by an antibody is called an epitope. The interaction of epitope and BCR initiates the B cell response. In the effector phase, binding of secreted antibodies masks the antigens and thereby triggers various effector mechanisms that finally result in the elimination of the antigen. Antibodies can also neutralize targets directly, for example, by binding to a part of a pathogen needed to cause an infection.

1.3.2 Combinatorial diversity - V(D)J recombination

The large and diverse population of antibodies is initially generated in the bone marrow by random combinations of a set of gene segments. These segments are called variable (V), diversity (D) and joining (J) segments or shortly, germline sequence, as they stem from the DNA germline of a cell. V, D and J segments are found in Ig heavy chains, but only V and J segments are found in Ig light chains. For example, human heavy chains may be built by one of 65 V, 27 D and 6 J gene segments, respectively. Additionally, insertions and deletions of nucleotides among the gene segments are generated. The rearrangement of these segments provides a huge diverse population of antibodies with their unique antigen binding site (paratope). The rearrangement process is called somatic DNA recombination of immunoglobulins, also known as V(D)J recombination. After a B cell produces a functional immunoglobulin gene by V(D)J recombination, it cannot express any other

variable region - this process is known as allelic exclusion. Therefore, each B cell can produce antibodies containing only one kind of variable chain. Because formation of the B cell repertoire in the bone marrow is antigen independent, it is also referred to as the pre-immune repertoire.

Summarized, the pre-immune antibody diversity is mainly determined by three factors [50]:

1. Germline variation of VDJ gene segments,

2. Recombination of VDJ gene segments and different pairing of the heavy and light chains,

3. Extra junctional diversity generated by insertion and deletion of nucleotides among the VDJ segments.

After antigen contact, further diversity is induced through point mutations. This process, taking place in germinal centers, is called affinity maturation and leads to an efficient recognition of the invading antigen by the antibody (section 1.3.3).

Clonal relatedness

The question if B cells stem from the same clone has implications on several immunological questions. Immunologists assume that if two B cells have identical CDR3 sequences, then they must have originated from the same parent cell. This assumption is due to the enormous combinatorial diversity during the rearrangement process, especially due to the number of modifications (deletions and insertions) between the V(D)J gene segments.

Usually, investigated sequences contain additional somatic hypermutations, which complicates the identification of clonal relatedness. The main approach for identifying clonal relatedness is done via junction analysis [95], which identifies the original (D)J genes. Since the diversity introduced by rearrangement-inserts is much broader than can be covered by simply identifying the original genes, the problem usually remains unsolved, even to this day. For light chains it is assumed that a plausible determination of clonal relatedness is not even possible [115]. Because of the involvement of only two genes in the rearrangement, the probability for erroneously judging two sequences to stem from the same clone is too high.

1.3.3 The germinal center reaction

The germinal center

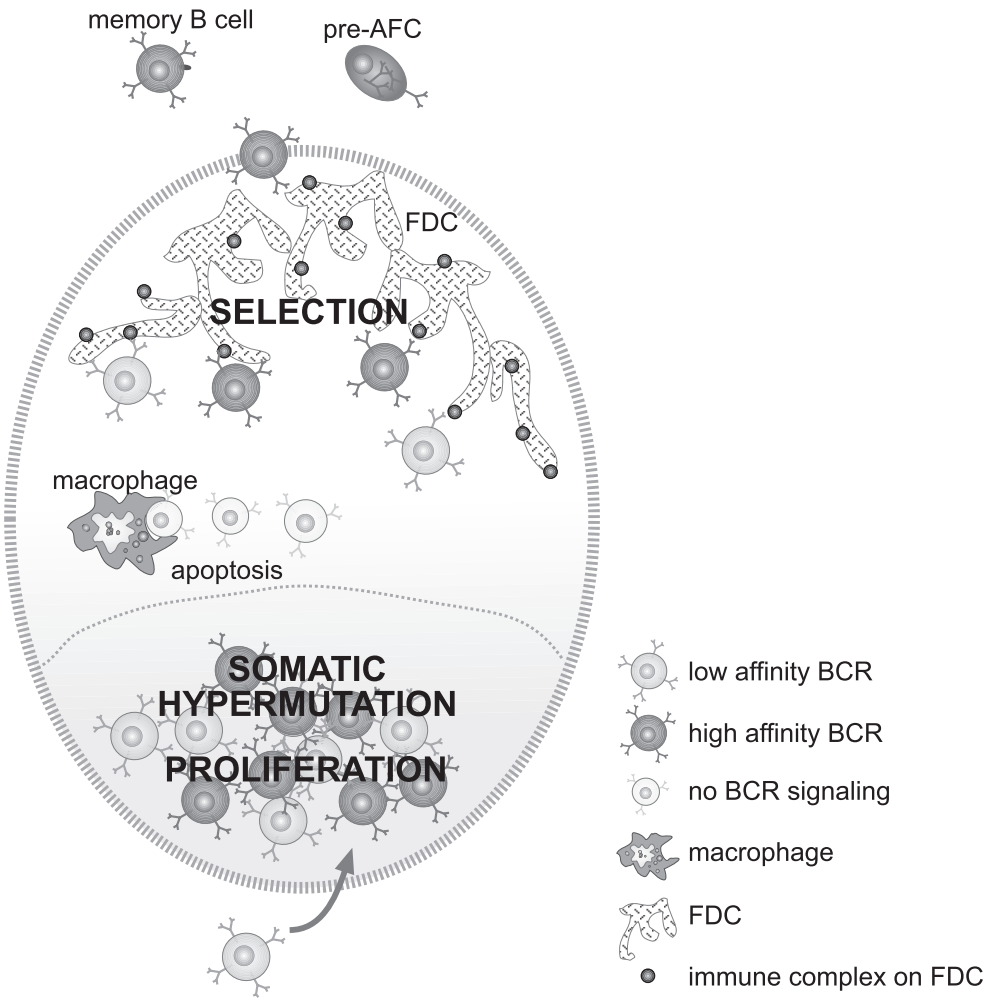
The term germinal center (GC) was introduced by Walther Flemming - a pioneer in the field of cytogenetics - more than 120 years ago. While studying cell division in lymph nodes, he observed strong proliferation of B cells within the follicles and suggested that these sites constitute the origin of lymphocyte generation or germination, respectively [40]. Flemming assigned the term germinal center to these structures according to their supposed function. The term persists until today, although we now know that not Flemming's GC but bone marrow and thymus are the primary lymphoid organs and sites of lymphocyte differentiation. Nevertheless, Flemming's interpretation was not entirely incorrect, because GC are indeed the source of generation of unique B cells (GC B cells) featuring high affinity B cell receptors (Figure 1.7). In addition, the formation of B cell memory, as well as immunoglobulin isotype class-switching are likewise associated with the GC reaction.

Affinity maturation

Central to the adaptive immune response is the process of affinity maturation that promotes efficient defense against pathogens by generating high affinity antibody secreting cells and memory B cells. This process was first assigned to the locale of the GC (Figure 1.7) about 20 years ago [8, 71, 97, 105]. GCs represent specialized microenvironments formed within secondary lymphoid tissues like the spleen or lymph nodes during T cell dependent immune responses [9, 81, 87, 89].

B cells that show moderate affinity to the pathogen are preselected. Within GCs, these B cells undergo monoclonal expansion for about three days and thereafter extensive alteration due to somatic hypermutation (SHM) of the genes encoding the V region of their BCRs [29, 100]. SHM involves random substitutions, deletions and insertions.

After diversification by SHM, B cells are subject to a receptor-mediated selection step that is based upon their ability to both bind an antigen and compete for interaction with accessory cells such as T cells and follicular dendritic cells [82, 104]. B cells that have acquired mutations beneficial with respect to antigen binding are positively selected and eventually differentiate into antibody secreting plasma cells or memory B cells that will be activated in subsequent contacts with the same antigen. Some of them might undergo further rounds of proliferation, mutation and selection [2, 99]. In contrast, B cells carrying disadvantageous or deleterious mutations are negatively selected and promptly eliminated by apoptosis. Allen *et al.* could show that



Source: Dissertation of Nicole Wittenbrink, HU Berlin, 2007

Figure 1.7: Illustration of a GC reaction. Activated B cells enter the GC and undergo affinity maturation. During proliferation B cells are subjected to somatic hypermutation. B cells that obtain high affinity are selected due to engagement of their BCRs and immune complexes deposited on follicular dendritic cells (FDC). B cells featuring low affinity become apoptotic and are immediately engulfed and degraded by macrophages. Positively selected B cells interact with T cells, resulting in final differentiation to either antibody-forming cells (AFC) or memory cells that exit the GC.

beneficial mutations enhance binding affinity within the range of one order of magnitude [3].

1.3.4 Analysis of B cell receptor sequences

Distribution of somatic hypermutations

Analyses of primary BCR sequences from GC derived B cells first revealed that somatic mutations are not distributed randomly throughout V regions. This was initially considered as a direct outcome of selection [122, 123]. However, mutations in sequences that are not influenced by the cellular selection process are also non-randomly distributed, demonstrating that the mutation mechanism itself has an intrinsic targeting bias [10, 33, 49]. The mutation pattern was shown to be affected by the local DNA microsequence context. This led to the identification of several preferentially targeted hotspot motifs such as RGWY (R=purine, Y=pyrimidine, W=A/T) and TAA [10, 114]. The intrinsic targeting bias is not exceedingly high, with values in hot spots not reaching higher than about three times the average mutability. These observations were generalized by the work of Shapiro *et al.*, who found a consistent hierarchy of mutability among all di- and trinucleotide sequences [118, 127].

Mouse models

The widespread use of classical models including hapten-carrier immunization in mice led to fine characterization of the appendant B cell responses. One typical feature of the immune response to certain haptens is the recurrent expression of a particular variable region gene [50, 59] as illustrated in Figure 1.8. This characteristic feature enables the study of complex processes taking place in germinal centers. Hence, it reveals some ideas about the features of B cells that underwent selection in GCs including the tendency of the replacement to silent mutation ratio (R/S) to be higher in CDRs than in framework regions (FRs). Also, the presence of certain recurrent key mutations [8, 24] that increase affinity by up to an order of magnitude was observed. However, since the observed mutation patterns of affinity matured antibodies are (besides selection) biased by the intrinsic mutability of their V regions, the debate regarding the significance of the supposed features associated with antigenic selection continues [12, 32, 54, 84].

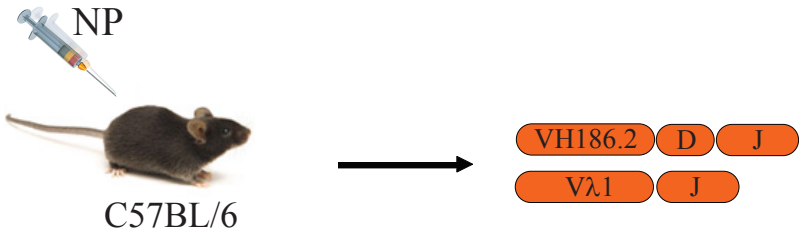


Figure 1.8: NP-immunization scheme of C57BL/6 mice. The antigen (4-hydroxy-3-nitrophenyl)acetyl (NP) is coupled to a non-affecting protein (chicken γ-globulin (CGG)) in order to trigger an immune response. The beneficial effect on this setup is that mostly the same germline V genes ($VH_{186.2}$ for the heavy chain, $VL_{\lambda 1}$ for the light chain) are selected for affinity maturation in germinal centers, which allows statistical analysis of the mutations induced by somatic hypermutation.

1.3.5 Nucleotide sequence databases

Theoretical investigations on the characteristic features of the germinal center reaction like the mutation pattern of BCR sequences require a large number of nucleotide sequences to make reliable conclusions. Gaining these nucleotide sequences experimentally is expensive and time-consuming. Alternatively, nucleotide sequence databases provide the opportunity to retrieve sequence data. However, sequence databases are not standardized in terms of format and organization of data, which complicates finding the desired information completely. For example the clonal relatedness of BCR sequences is, in most cases, completely unclear.

Some of the most important databases are:

- The International Nucleotide Sequence Databases (INSD) have been developed and maintained collaboratively between DDBJ, EMBL, and GenBank (http://insdc.org).
- The *EMBL* Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications (http://www.ebi.ac.uk/embl).
- *VBASE2* is an integrative database of germline variable genes from the immunoglobulin loci of human and mouse. All variable gene sequences are extracted from the EMBL-Bank (http://vbase2.org).
- *IMGT*, the international ImMunoGeneTics information system contains information similar to VBASE2, such as germline genes of human

and mouse. Additionally, it offers alignment, identification of genes and further useful tools. For example IMGT/V-QUEST (V-QUEry and STandardization) is an integrated alignment tool for the Ig nucleotide sequences. IMGT/V-QUEST compares rearranged Ig variable sequences with its own reference sets (http://www.imgt.org).

1.4 Measuring protein-protein interactions

Several methods to measure protein-protein interactions and binding affinities are commonly used, such as enzyme linked immunosorbent assay (ELISA) or surface plasmon resonance (SPR). But such methods are rarely suitable for high-throughput measurements in contrast to protein- and peptide arrays on cellulose membranes or glass slides [41, 5], that allow higher density of probes and the parallel measurement of a multitude of peptide-protein interactions. Direct comparison of antigen arrays with ELISA demonstrated that antigen arrays were consistently four- to eight-fold more sensitive for detecting autoantibody binding to five autoantigens tested [113].

As demonstrated by microarray experiments, the array format is a robust, reliable and well-established method for high-throughput analysis for both gene expression and protein-protein interactions in a single experiment [34, 36, 86]. The advantage of peptide opposed to protein libraries is manifold: firstly, peptides have greater stability, which allows experiments at room temperature, and the storage of arrays for months. Secondly, peptides can be synthesized quickly, cheaply and with high purity. Furthermore, problems due to misfolding are avoided.

1.4.1 The SPOT synthesis technique

The SPOT technology is a widely used technology for high-throughput investigations of peptide/protein interactions. It allows the experimental analyses of large peptide libraries for the purpose of studying binding properties of peptides to proteins and the benefit of studying asymmetric protein-protein interactions: a domain of partner A acts as a binding partner for a linear peptide in partner B. In fact, a fairly large set of cellular protein interactions are mediated by families of relatively small protein interaction domains (PID), such as SH2, SH3, GYF, WW, or PDZ, which act as receptors accommodating short peptides in their binding pockets. The SPOT technology is an ideal tool for preparing synthetic peptide arrays, because it is an easy-to-use, robust and inexpensive method. Furthermore, the technique and many of its applications have been reviewed extensively [41, 108, 142].

In principle, signal intensities - the output of this technique - can be used to roughly distinguish between different affinities. This was shown for peptide-antibody interactions [73, 135]. However, the most important application of the SPOT synthesis technique is to differentiate qualitatively between binding affinities of peptides to defined protein targets within one array. More details on the setup can be found in the Materials and Methods section 3.1.1 and are illustrated in Figure 3.1.

CB4-1 antibody

The murine monoclonal antibody (mAb) CB4-1 raised against p24 (HIV-1) recognizes a linear epitope of the HIV-1 capsid protein [51, 56]. Additionally, CB4-1 exhibits cross-reactive binding to epitope-homologous peptides and polyspecific reactions to epitope nonhomologous peptides [55, 72].

The recognition profile of mAb CB4-1 is a well-established experimental system that has already been used to evaluate a variety of structurally different types of peptide arrays prepared by the SPOT technology. This includes scans of overlapping peptides [142], randomly generated peptide libraries [109], transition pathway libraries that reveal evolutionary connections between different peptides with similar activities [55], and peptide arrays to elucidate the cross-reactivity of mAb CB4-1 [72].

1.4.2 Covering sequence space - Epitope search

The identification of peptides that bind to antibodies is an important step in characterizing antibody specificity in order to study molecular recognition occurring during humoral immune responses and to investigate cross-reactivity potentially implicated in autoimmune diseases. In addition, many processes using antibodies as research tools, diagnostics, reagents, or therapeutics require more detailed information about their interaction with peptide antigens.

Identification of antibody binding peptides may be based on the primary structure of the antigens used to raise the antibodies (knowledge- or sequence-based approach). Peptide scan is the term for scanning the entire sequence of the antigen with overlapping peptides, usually not longer than 15 amino acids, which are probed for binding to the respective antibody. The sequence common to the interacting peptides is the epitope [46]. To map linear epitopes [6], peptide scans are an easy and straightforward approach. In linear epitopes the residues that are effectively in contact with the antibody are located within only one stretch of the protein sequence usually not exceeding 12 amino acids [30]. Characteristically, linear epitopes identi-

fied using peptide scans have affinities to the antibody that are only slightly lower than the affinity of the antibody to the entire antigen. In principle, all multiple peptide synthesis strategies are well suited to prepare peptide scans. However, the SPOT synthesis technique [41, 142] has emerged as the most practical and economical since only small amounts of amino acids and reagents are required in the synthesis. The mapping of discontinuous epitopes [6, 48] is a far more challenging task. In these binding sites the key residues are distributed over two or more binding regions, usually located far apart in the primary structure, which upon folding are brought together on the protein surface to form a composite epitope. Even if the complete epitope elicits a high affinity interaction, peptides covering only one binding region, as synthesized in a peptide scan, generally have very low affinities that often cannot be measured in normal ELISA or SPR experiments.

If the antigen is not known, or if potential cross-reactivities of, for example, autoantibodies have to be investigated, *de novo* approaches are required to identify possible peptide antigens. There have been several advances with arrays of peptides and peptide mixtures prepared by SPOT synthesis [72, 109]. The most complex SPOT library described so far is one of the type XXXX[3O3X]XXXX. The internal core [3O3X] is an abbreviation for three defined and three randomized positions arranged in all possible combinations, e.g. XXXX[$O_1O_2O_3$XXX]XXXX; XXXX[O_1O_2XO_3XX]XXXX and so on [72]. This library comprised $68,000$ peptides and has been used to identify not only antibody epitopes but also other peptides that bind to the paratope of the antibody in a completely different mode (referred to as mimotopes). With a different but much simpler approach by just using a random library of 5520 peptides, Reineke [109] yielded similar results. He showed that his approach is sufficient to rapidly and economically select peptidic antibody epitopes and mimotopes.

Chapter 2

Objectives

In this work, several aspects of peptides, epitopes, antibodies and their binding affinity to each other depending on mutations within peptide or DNA sequences are to be investigated. Two main scientific goals for this work are formulated: First, to elucidate the relationship of amino acids in terms of binding affinity changes. Second, to provide new insights into the evolution of BCRs during the immune response. The dataset for the estimation of amino acid relationships shall be the substitution analysis data prepared by Liying Dong in her thesis by using the SPOT synthesis technique [30]. This dataset promises to be a great starting point.

In detail, the following tasks are to be executed within the scope of this work:

1. Analyze the reliability of the SPOT synthesis technique in detail.
2. Generate a substitution matrix based on the data of Liying Dong.
3. Analyze coherences between amino acids.
4. Establish reduced sets of amino acids and check the applicability for epitope search.
5. Analyze the functionality of the epitope search approach of the immune response with the help of the substitution matrix.
6. Develop a method to identify the most important mutations of a specific immune response.

For the widespread use of the SPOT synthesis technique it is expected that this work supports quantitative evaluation of signal intensities and gives assistance for library design. The substitution matrix might provide a new

basis for amino acid similarities, which might be useful especially for antibody binding investigations. Additionally, the reduced sets of amino acids should be a good partner for epitope searching. Finally, the new method for the identification of important mutations in these kinds of analyses is an objective means to gain new insights into the topic.

Chapter 3

Materials and Methods

3.1 Protein-peptide interaction measurements

3.1.1 Synthesis of the cellulose membrane-bound peptide arrays

Cellulose-bound peptide arrays were semi-automatically prepared according to standard SPOT synthesis protocols using a SPOT synthesizer (Intavis, Koeln, Germany) as described in detail [141] and illustrated in Figure 3.1. The peptides were synthesized on amino-functionalized cellulose membranes of the ester type [142] prepared by modifying a cellulose paper (Whatman 50, Whatman, Maidstone, UK) with Fmoc-β-alanine as the first spacer residue (3.1a). Loading of the amino-functionalized cellulose membrane was determined as described before [141]. In the second step of coupling (3.1b) the anchor position, a mixture of different ratios of a 0.3 M solution of Fmoc-β-alanine-OPfp and a 0.3 M solution of N-acetyl-β-alanine-OPfp in dimethylsulfoxide (DMSO) was used (100%, 50%, 25%, 10% Fmoc-β-alanine-OPfp) leading to membranes with a 100%, 50%, 25% and 10% amino-functionalization quotient (FQ) at the spot-positions [73]. Residual amino functions between the spots are capped by acetylation (3.1c). Cleavage of the Fmoc group was done with 20% piperidine in dimethylformamide (DMF). The cellulose-bound peptide arrays were assembled on these membranes (3.1d) by using 0.3 M solutions of Fmoc-amino acid-OPfp in NMP (in case of Ser and Thr the ODNp-esters were used). Side-chain protection of the used Fmoc-amino acids was as follows: Glu, Asp (OtBu); Ser, Thr, Tyr (tBu); His, Lys, Trp (Boc); Asn, Gln, Cys (Trt); Arg (Pbf). After the last coupling step, the acid-labile protection groups of the amino acid side chains were cleaved using 90% TFA (trifluoro acetic acid) for 30 minutes and

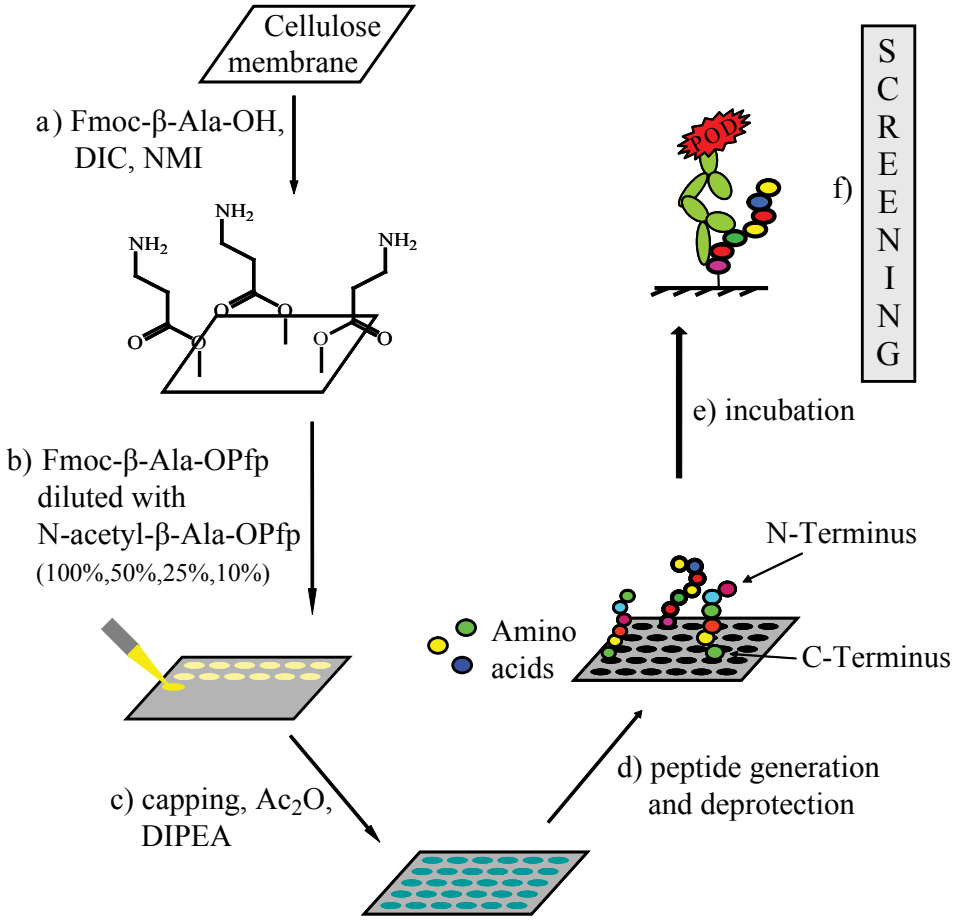


Figure 3.1: Scheme of the SPOT synthesis technique. a) The cellulose membrane is modified with Fmoc-β-alanine. b) The membrane is loaded with different ratios of Fmoc-β-alanine-OPfp and N-acetyl-β-alanine-OPfp leading to different coupling opportunities for amino acids. c) The spots are separated by capping possible residual amino functions between them. d) Generation of peptides by pipet-robots. e) Incubation with a protein/antibody. f) Measurement of signal intensities.

thereafter 50% TFA for 3 hours.

Chemistry files containing the information about the composition of the peptides were generated using the custom-made software "LISA", which can be obtained by contacting the author. The chemistry files serve as input for the semi-automated synthesis done by pipet-robots that add single amino acids to the peptides. After adding of an amino acid to each peptide, the membrane is moved from the robot to a washtub for manual washing according to the protocol. Prior to the next pipet-step the membrane was again correctly fixed on the robot.

For synthesis quality control of the peptide library, a representative selection of peptides was prepared (spot size 0.25 cm^2) and cleaved by ammonia vapor in the dry state [136, 141]. Subsequently, identity was verified by matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) (LaserTec BenchTop II, Applied Biosystems, Forser City, CA, USA) and peptide quality was tested by analytical reversed phase HPLC (Waters 600, Waters, Eschborn, Germany) on a C-18 column (Vydac, Hesparia, CA, USA).

According to the general protocol a standard peptide array (dimension 6 cm x 8 cm) was synthesized on cellulose membranes containing 37 different peptides (Table 3.1, entries $2-38$) with an average of 13 repeats of each peptide and additionally 120 times the peptide GATPEDLNQKLAGN (Table 3.1, entry 1) serving as high affinity reference for statistical purposes.

The peptides were equally distributed on the array resulting in a cellulose membrane that contains 607 spots. Such an array was generated on membranes with four different FQ (100%, 50%, 25%, 10%) leading to the standard cellulose membranes SCM-100, SCM-50, SCM-25, SCM-10. The membrane SCM-50 was prepared three times and each replica was incubated with a different antibody concentration. A membrane with the same FQ and the same peptides as SCM-50 was prepared, but with about the three- to four-fold number of each peptide resulting in the membrane CM_{2090}-50 with 2090 spots. Furthermore, the cellulose membrane CM_{6000}-50 ($FQ = 50\%$) was prepared using six different peptides from Table 3.1 (entries: 1, 6, 14, 20, 26, 36), but with each peptide equally distributed as 1000 reiterations on the array (size: 18 cm x 24.4 cm). The peptides cover the affinity spectrum from $pK_{dis} = 9.0$ to $pK_{dis} = 3.5$.

3.1.2 Binding studies on cellulose membrane-bound peptides

All incubation and washing steps were carried out under gentle shaking and at room temperature, unless stated otherwise. After washing the membrane

No.	Sequence	pK_{dis}	No.	Sequence	pK_{dis}
1	GATPEDLNQKLAGN[a]	9.0	20	RDFDKEWNLIEQNS[a]	5.7
2	GATPQDLkTML[b]	9.0	21	LELIQDLNQKLQDGF[d]	5.7
3	GATPEDLNQKL[c]	8.2	22	TTMEWFRITDGARIM[d]	5.7
4	sATPwDLkTsl[#]	7.7	23	sAGPwDLkssl[b]	5.3
5	FATPEDLNQKL[c]	7.7	24	LEMKQDLNQKLQDGF[d]	5.3
6	GATPQDLNTML[#]	7.3	25	LEMKQDLNQMLQDGF[d]	5.2
7	GLKEWGGARIT[#]	7.1	26	DYFDLTQDNIITRRL[d]	5.0
8	DATPEDLNAKL[#]	7.0	27	LEMKQDLNIKLQDGF[d]	5.0
9	DATPEDLNARL[#]	7.0	28	efslkGpllqwrsG[a]	4.7
10	GATPQDLkTMl[#]	6.9	29	DYFDLTQDNIITRRN[d]	4.7
11	GATPEDLNAKL[#]	6.8	30	LEMKQDLNPMLQDGF[d]	4.7
12	FDKEWGGIRIT[c]	6.8	31	sAGdwwLkssl[b]	4.5
13	DALYEWGGARI[c]	6.7	32	saGdwwGkssl[b]	4.4
14	GLYEWGGARITNTD[a]	6.7	33	sAGdwDLkssl[b]	4.3
15	sATPwDLkTMl[#]	6.6	34	saGdwwLkssl[b]	4.2
16	DALPEWGGARI[#]	6.2	35	ITDGARIM[d]	4.0
17	GATPwDLkTMl[#]	6.1	36	DYFDLTQDNIIERRN[d]	4.0
18	sATPwDLkssl[b]	6.0	37	LEMKQDLNIMLQDGF[d]	4.0
19	EAWVLEGAMILWKTD[c]	6.0	38	EAWVLRGAMILWKTD[d]	3.5

= BIAcore studies; a = [72]; b = [109]; c = [55]; d = [135].

Table 3.1: Dissociation constants obtained for mAb CB4-1/peptide complexes.

with ethanol once (10 min) and three times for 10 min with Tween-Tris buffered saline (T-TBS: 50 mM Tris-(hydroxymethyl)-aminomethane, 137 mM NaCl, 2.7 mM KCl, adjusted to pH 8 with HCl/0.05% Tween 20), the membrane bound peptide arrays were blocked (3 h) with blocking buffer, i.e., blocking reagent (CRB, Northwich, UK) diluted 1:10 in T-TBS containing 5% (w/v) sucrose, and then washed with T-TBS (1x10 min). Subsequently, the peptide arrays were incubated with the anti-p24 (HIV-1) monoclonal antibody (mAb) CB4-1 (mouse IgG2a/k) [51, 56] at different concentrations in T-TBS blocking buffer for 16 h at 4° C (Figure 3.1e). The following concentrations of mAb CB4-1 were used: 0.1μg/ml, 1μg/ml and 10μg/ml, whereas the concentration of 1μg/ml was used with the membranes *SCM*-10, *SCM*-25, *SCM*-50 and *SCM*-100. After washing three times for 10 min with T-TBS, the second anti-mouse IgG peroxidase-labeled antibody (Sigma, Deisenhofen, Germany) was added at a final concentration of 1μg/ml in T-TBS blocking buffer for 2 h, followed by washing three times with T-TBS. Analysis and quantification of peptide-bound mAb CB4-1 (3.1f) was carried out using a chemiluminescence substrate and a Lumi-Imager^{TM} (Roche Diagnostics, Basel, Switzerland).

3.1.3 Measurement of spot signal intensities

Analysis and quantification of spot signal intensities (SI) were executed with the software Genespotter® (MicroDiscovery GmbH, Berlin, Germany). Genespotter® has a fully automatic grid finding routine resulting in reproducible signal intensities. The spot signal is calculated from a circular region around the spot center detected on the image. The local background signal for each spot is determined with a safety margin to this circular region.

3.1.4 Substitution analysis

Substitution analysis stands for the exchange of amino acids on each position of a peptide. The exchange is complete meaning that a substitution into each of the natural occurring 20 amino acids is applied (see illustration in Figure 3.2). As a result one gets complete information about binding affinities in the direct neighborhood (Hamming distance one) of the peptide.

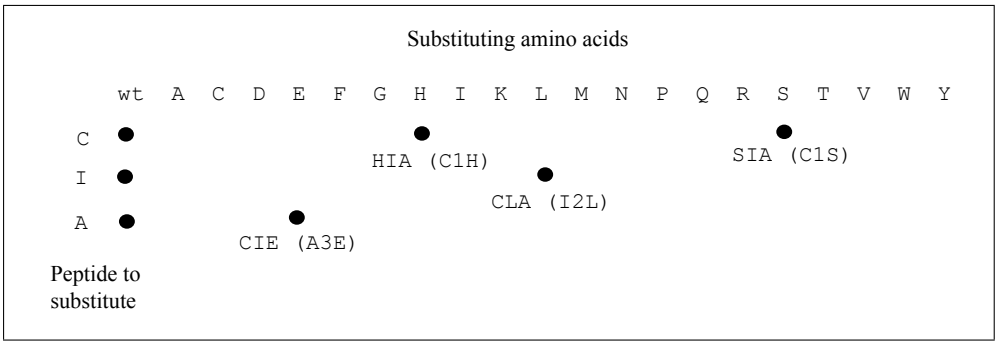


Figure 3.2: Scheme of the substitution analysis. Substitution analysis stands for the exchange of amino acids on each position of a peptide (here: CIA) against each naturally occurring amino acid. The first column represents the wildtype, which is the peptide itself. Any further column is a single substitution of the amino acid in the according row into an amino acid of the according column (Ala, Cys, Asp, ...). Each black spot indicates binding of the incubated ligand on the according peptide.

3.1.5 Standard solid phase peptide synthesis

Soluble peptides were synthesized (50 μmol scale) as amides on a multiple synthesizer AMS 422 (Abimed, Langenfeld, Germany) according to the standard Fmoc machine protocol using TentaGel S RAM resin (Rapp Polymere, Tübingen, Germany) and PyBOP activation. All peptides were analyzed by reversed phase HPLC on a Vydac C18 column using a linear gradient of 5-60% acetonitrile/water (0.05% TFA) for 20 min at 1.2 ml/min flow-rate

(detection at 214 nm) and by MALDI-MS using α-cyano-4-hydroxy-cinnamic acid as matrix and purified to >95% purity by preparative HPLC on a Vydac C18 column, if necessary.

3.1.6 SPR-Measurement

The affinity of peptides in solution to the mAb CB4-1 was measured in terms of K_{dis} by monitoring adsorption-dependent surface plasmon oscillations using the BIACORE®X system (Biacore, Uppsala, Sweden). Both the mAb CB4-1 and anti-GST monoclonal antibody (control antibody) were immobilized on a dextran-coated gold surface in separate flow cells of a CM5 sensor chip (Biacore AB, Uppsala, Sweden) using the amine coupling procedure (5 μl/min; activation with 1:1 EDC/NHS, 7 min; immobilization with 50 μg/ml antibody, 5-75 μl multiple injections; deactivation of excess groups with 1 M ethanol amine-HCl pH 8.5, 7 min). The amount of covalently coupled antibodies corresponded to a signal increase of approximately 5000 resonance units (RU) for both antibodies (flow cell 1: control antibody; flow cell 2: mAb CB4-1). All binding experiments were performed at 20° C with a flow rate of 15 μl/min (injection volume 10 μl). Peptides were used at various concentrations between 20 pM and 200 mM in HBS (10 mM Hepes with 0.15 M NaCl, 3.4 mM EDTA, and 0.005% surfactant P20, pH 7.4). Complete regeneration was obtained after dissociation without regeneration buffer in most cases. Different regeneration procedures (BIAcoreX manual) were required in some cases. Transformation of data and analysis were performed with the BIA-evaluation software, version 3.0. The control sensorgram (flow cell 1) was subtracted from the sensorgrams obtained with flow cell 2. The steady-state values of the binding equilibrium were plotted versus the different peptide concentrations and fitted using the implemented steady-state evaluation, resulting in the K_{dis} for each model peptide (Table 3.1).

3.2 Collection and analysis of antibody sequences

3.2.1 Extraction and selection of $VH_{186.2}$ sequences

All $VH_{186.2}$ sequences used in this thesis were extracted from the EMBL Nucleotide Sequence Database [23] and originate from previous investigations of the primary anti-NP-chicken γ-globulin (CGG) response in C57BL/6 mice [19, 42, 43, 44, 57, 58, 60, 64, 89, 121, 129, 130, 131, 132, 144]. The collected

$VH_{186.2}$ sequences were isolated from various tissues including bone marrow (BM), spleen, lymph nodes and nasal-associated lymphoid tissue. The latter three are subsequently referred to as lymphatic tissue (LT). Moreover, $VH_{186.2}$ sequences were experimentally obtained at different time points after primary immunization, starting from day 6 up to day 140. A detailed survey of the extracted $VH_{186.2}$ sequences is given in Table 3.2. Each extracted sequence was initially rechecked against both the IMGT [76] and VBASE2 [111] database of Ig genes in order to ensure that the closest alignment results in the $VH_{186.2}$ gene. Although there are investigations concerning the reliability of the identification of human immunoglobulin genes [75, 139], the identification of the murine $VH_{186.2}$ gene is considered univocal and noncritical since it is very well known and its nearest neighbor (IGHV1-53*04) has a Hamming distance of 10 according to the IMGT database. After sequences were divided into FRs and CDRs according to the IMGT unique numbering system [77], they were compiled into one integrated aligned library using a purpose-built custom-made Java application. To avoid overestimation of the frequency distribution of mutations due to clonal redundancy, the additional selection criteria were applied for sequences stemming from the same mouse:

- clonal independence of $VH_{186.2}$ sequences (as specified below)
- independence of mutations among potentially clonally related $VH_{186.2}$ sequences, indicated by a unique mutation pattern. In the case of one more identical mutation between FR1 and FR3, only the $VH_{186.2}$ sequence showing the highest number of mutations was included in the dataset.

3.2.2 Assessment of clonal independence

$VH_{186.2}$ sequences originating from the same publication or, in case this information was available, from the same individual mouse, were checked regarding their clonal relatedness. The sequences were compared pairwise, and if a pair was found to be potentially clonally related according to the algorithm described below, the sequence with the smaller number of mutations was excluded from the final dataset. To estimate the clonal relatedness of two sequences (Figure 3.3), the maximal number of mutations expected to be found in the CDR3 of every single sequence ($M_{3,max}$) was calculated. This calculation is based on the assumption that M_3, the number of mutations expected in the CDR3, depends directly on the empiric mutation probability q found on the appending V gene, and on the nucleotide length of the CDR3 (S_3).

Day[a]	**Tissue**[b]	**VH**$^{c}_{186.2}$	**[average]**[d]	**gl [%]**[e]	**W34L [%]**[f]
Early phase					
6	LT	9	1.33 (0.44)	67	0 (0)
7	LT	69	1.94 (0.51)	74	0 (0)
8	LT	20	2.73 (2.05)	25	0 (0)
9	LT	70	2.77 (2.37)	14	12 (10)
10	LT	56	2.93 (2.25)	23	9 (7)
$\sum$		224	2.68 (1.66)	38	8 (5)
Peak phase					
11	LT	82	2.49 (1.43)	43	11 (6)
12	LT	22	4.33 (2.95)	32	33 (23)
	BM	7	4.25 (2.43)	43	0 (0)
13	LT	115	2.11 (1.39)	34	13 (9)
14	LT	10	4.88 (3.90)	20	38 (30)
	BM	1	9.00 (9.00)	0	0 (0)
15	LT	38	2.67 (1.47)	45	10 (5)
20	LT	17	6.35 (6.35)	0	47 (47)
21	LT	29	2.44 (1.52)	38	22 (14)
$\sum$		321	2.97 (1.92)	36	18 (12)
Late phase					
30	LT	21	4.69 (2.90)	38	8 (5)
40	LT	28	6.35 (5.21)	18	13 (11)
	NS	5	1.75 (1.40)	20	0 (0)
42	LT	7	6.25 (3.57)	43	25 (14)
	BM	5	7.00 (5.60)	20	0 (0)
46	BM	19	3.77 (2.58)	32	23 (16)
60	LT	10	4.30 (4.30)	0	10 (10)
63	LT	5	11.00 (6.60)	40	33 (20)
	BM	8	6.29 (5.50)	13	43 (38)
69	BM	11	5.80 (5.27)	9	40 (36)
80	LT	24	5.43 (3.17)	42	14 (8)
85	NS	18	3.25 (2.17)	33	25 (17)
119	BM	16	4.00 (3.25)	19	31 (25)
120	NS	29	4.75 (3.28)	31	50 (34)
140	LT	2	9.00 (9.00)	0	0 (0)
$\sum$		208	5.09 (3.72)	27	24 (17)
Undefined time points					
-	LT	28	1.50 (0.11)	93	0 (0)
$\sum$		781	3.53 (2.26)	36	17 (11)

[a] $VH_{186.2}$ sequences were obtained at the indicated time points after primary NP-CGG challenge of C57BL/6 mice.
[b]Tissue origin of collected $VH_{186.2}$ sequences.
[c]Number of collected $VH_{186.2}$ sequences.
[d]Average mutation frequency of $VH_{186.2}$ sequences not considering germline sequences; values in brackets indicate average mutation frequency when all sequences are included.
[e]Proportion of germline (gl) $VH_{186.2}$ sequences.
[f]Proportion of $VH_{186.2}$ sequences bearing the key mutation (W34L) not considering germline sequences; values in brackets indicate the proportion of W34L mutations when all sequences are included.

Table 3.2: Survey of collected $VH_{186.2}$ chain sequences

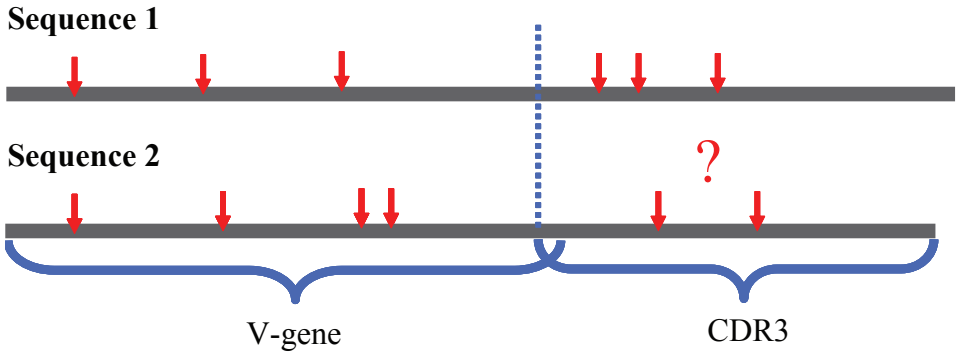


Figure 3.3: Estimation of clonal independence. The estimation of clonal independence of two sequences is done by measuring the Levenshtein distance of their CDR3 regions, which estimates the number of somatic hypermutations therein. If the two sequences stem from different clones, this distance must be "large". The number of somatic hypermutations observed in the V genes of the two sequences are used to estimate the minimal distance between the two CDR3s that is expected for reliable identification of independency. This is done in the Equations (3.1 - 3.4).

To be on the safe side, all positions in a CDR3 are assumed to be hotspots (Equations (3.1, 3.2)).

The estimation of clonal unrelatedness was done using the cumulative distribution function (F) (Equation (3.4), Figure 3.4) of the binomial probabilities of the expected numbers of mutations within the CDR3 (Equation (3.3)), in which the expected maximal number of mutations ($M_{3,max}$) was defined as the largest M_3 with $F(S_3, M_3) \leq 0.95$. A pair of sequences was rated as clonally unrelated if the sum of their expected maximal numbers of mutations in the CDR3, $M_{3,max}$, is smaller than the real Levenshtein distance between their respective CDR3s. The Levenshtein distance between two sequences is given by the minimum number of operations (mutations, insertions or deletions) needed to transform one sequence into the other.

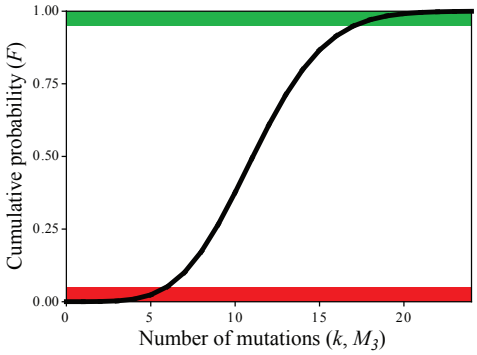


Figure 3.4: Example of the cumulative binomial distribution. On the x-axis the number of mutations is shown. The y-axis gives the associated cumulative probability. The 0.95 confidence interval (green and red areas) indicates the maximal/minimal number of expected mutations for an exemplary total quantity (S_3, N) and probability (q_{max}, $p^0_{m,i}$) as specified in Equations (3.4, 3.10).

Equations used for estimation of clonal independency

In Equation (3.1), q is the (empiric) mutation frequency of the V gene of a $VH_{186.2}$ sequence as calculated by dividing the number of observed point mutations within this V gene (M_V) by its nucleotide length ($S = 288$).

$$q = \frac{M_V}{S} \tag{3.1}$$

In Equation (3.2), q_{max} is the maximal mutation frequency of the V gene of a $VH_{186.2}$ sequence as calculated by multiplying its observed mutation frequency by the maximal relative positional mutability observed in the dataset ($f_{i,max} = 2.81$). This is equivalent to assuming every position of the V gene to be a hotspot.

$$q_{max} = q * f_{i,max} \tag{3.2}$$

In Equation (3.3), P gives the binomial probability of a certain number of mutations within the CDR3 of a $VH_{186.2}$ sequence (M_3) as calculated regarding the nucleotide length of the CDR3 (S_3) and the maximal mutation frequency of the appendant V gene (q_{max}).

$$P(S_3, M_3, q_{max}) = \binom{S_3}{M_3} (q_{max})^{M_3} (1 - q_{max})^{S_3 - M_3} \tag{3.3}$$

In Equation (3.4), F is the cumulative distribution function of the binomial probabilities of the expected numbers of mutations within the CDR3 of a $VH_{186.2}$ sequence for $0 \leq j \leq M_3$.

$$F(S_3, M_3, q_{max}) = \sum_{j=0}^{M_3} P(S_3, j, q_{max}) \tag{3.4}$$

3.2.3 Predicting the relative frequency distributions of point mutations

The frequency distribution of mutations was predicted for the FR1 to FR3 region of the $VH_{186.2}$ chain (position 1 to 104). The prediction was solely based on the intrinsic mutability of the $VH_{186.2}$ chain, which includes the positional mutability of individual bases and the transition to transversion bias of mutations (Figure 1.4).

The positional mutability of each individual nucleotide at position i in the $VH_{186.2}$ sequence was calculated according to Shapiro *et al.* [119, 120], and comprises the average of trinucleotide mutability at position i and its immediate neighboring positions. The calculated positional mutability was divided by the average of the positional mutabilities for all positions, giving the relative positional mutability f_i. Thus, a value of 1 represents the average mutability for the $VH_{186.2}$ sequence. The relative codon positional mutability of a given amino acid in the $VH_{186.2}$ chain (f_c) was obtained by averaging the three relative positional mutabilities of its codon triplet.

The transition (purine-to-purine or pyrimidine-to-pyrimidine mutation, see Figure 1.4) bias of VH genes is established, and transitions were shown to occur twice as often as transversions [10, 137]. However, since each base can mutate into the three others, either by one transition or by two transversion substitutions, the mutation bias of individual bases p_m was factored in as $p_m = 0.667$ for transitions and $p_m = 0.167$ for transversions (Equation (3.5)).

The relative probability of a base in a given position of the $VH_{186.2}$ chain to mutate into one of the three others ($p_{m,i}$) was determined by multiplying its relative positional mutability f_i and the respective mutation bias p_m (Equation (3.6)). Correspondingly, the relative probability for a particular amino acid substitution (including synonymous and nonsynonymous substitutions) to occur in a given position of the $VH_{186.2}$ chain was calculated by averaging the relative probabilities of all mutations in the codon triplet that lead to the actual substitution.

Equations used for the statistical comparison

In Equation (3.5), p_m gives the probabilities for transitional and transversional point mutations (m) according to the well-established mutation bias [10, 137, 98].

$$p_m = \begin{cases} \frac{2}{3}, & \text{if } m \text{ transition} \\ \frac{1}{6}, & \text{if } m \text{ transversion} \end{cases} \tag{3.5}$$

In Equation (3.6), $p_{m,i}$ is the relative probability that a base in a given position i of the $VH_{186.2}$ sequence mutates into one of the three others, as calculated by multiplying its mutation bias p_m and its relative positional mutability f_i. The relative positional mutability of each individual base in the $VH_{186.2}$ chain (f_i) was obtained by dividing its respective positional mutability by the average of all positional mutabilities. The positional mutability was calculated according to Shapiro *et al.* [119, 120].

$$p_{m,i} = p_m * f_i \tag{3.6}$$

In Equation (3.7), p_0 is the empirical mutation probability as calculated by dividing the total number of observed point mutations ($M = 1859$) by the product of the number of extracted $VH_{186.2}$ sequences ($N = 781$) and the nucleotide length of the germline sequence of $VH_{186.2}$ without CDR3 ($S = 288$).

$$p_0 = \frac{M}{S * N} = \frac{1859}{288 * 781} = 0.008 \tag{3.7}$$

In Equation (3.8), $p^0_{m,i}$ is the predicted probability that a base in a given position i of the $VH_{186.2}$ chain mutates into one of the three others, as calculated by applying the empirical mutation probability p_0.

$$p^0_{m,i} = p_0 * p_{m,i} \tag{3.8}$$

In Equation (3.9), P gives the binomial probability of a given point mutation being identified k times within the dataset of all $VH_{186.2}$ sequences ($N = 781$) as calculated regarding its predicted probability ($p^0_{m,i}$).

$$P(N, k, p^0_{m,i}) = \binom{N}{k} (p^0_{m,i})^k (1 - p^0_{m,i})^{N-k} \tag{3.9}$$

In Equation (3.10), F is the cumulative distribution function of the binomial probabilities of a particular point mutation for $0 \leq j < k$.

$$F(N, k, p^0_{m,i}) = \sum_{j=0}^{k-1} P(N, j, p^0_{m,i}) \tag{3.10}$$

3.2.4 Observed frequency distributions of point mutations

The observed frequency distributions of point mutations in the investigated dataset were obtained by comprehensive sequence analysis of the FR1 to FR3 region (position 1 to 104) of collected $VH_{186.2}$ sequences at both the nucleotide and amino acid level. For identification of mutations, sequences were compared with the $VH_{186.2}$ germline sequence using the IMGT/V-QUEST integrated software program [47]. The empirical mutation rate of the $VH_{186.2}$ chain ($p_0 = 0.008$) was calculated by dividing the total number of observed point mutations ($M = 1859$), by the product of the number of extracted $VH_{186.2}$ sequences ($N = 781$) and the nucleotide length of the germline sequence of $VH_{186.2}$ without CDR3 ($S = 288$) (Equation (3.7)). The calculation of p_0 was performed including germline sequences.

3.2.5 Statistical identification of favored point mutations

Point mutations that occur more often (favored) than predicted were identified by statistical comparison of the predicted and observed frequency distributions of point mutations. The probability P of each particular point mutation $p_{m,i}$ being identified exactly k times at position i was estimated by factoring in the empirical mutation rate p_0 (Equation (3.8)) and using the binomial distribution (Equation (3.9), Figure 3.4). Favored point mutations were subsequently identified according to the 0.95 confidence interval of the cumulative distribution function F (Equation (3.10)). A point mutation occurring k times within the dataset was rated favored if $F(k) > 0.95$. Likewise, amino acid substitutions related to these point mutations were also rated favored.

Since this analysis involves a multiple testing problem, false discovery rate (FDR) control was applied according to Benjamini-Hochberg *et al.* [7]. As many tests as the potential number of point mutations within the FR1 to FR3 region of the $VH_{186.2}$ chain were simultaneously performed.

3.2.6 Localization of favored amino acid substitutions in the 3-D structure

For localization of favored amino acid substitutions in the $VH_{186.2}$ chain, the 3-D structure of a $VH_{186.2}/VL_{\lambda 1}$ Fv fragment complexed with an NP compound obtained from the Protein Data Bank [70] (PDB ID: 1a6v) was

visualized using the molecular visualization system PyMol [28]. Positions identified as featuring favored mutations were subsequently indicated in the 3-D structure.

3.2.7 Assessment of signatures of antigenic selection

For assessment of the signature of antigenic selection the expected and observed frequencies of replacement (R) and silent (S) mutations in the CDRs and FRs of the $VH_{186.2}$ gene were statistically compared applying the focused binomial test recently published by Hershberg *et al.* [54]. This test shows improved specificity compared to global binomial tests [84] and introduces the effects of microsequence specificity as well as the transition bias of somatic hypermutation (SHM) into the null model of mutation. Both the expected frequency under the null hypothesis of no selection (E_0) and the observed frequency are calculated as the fraction of R mutations within the region of interest (FR or CDR) and the overall number of mutations (R+S) within the whole sequence. Expected and observed frequencies were statistically compared using a two-tailed binomial test, with which sequences were rated according to showing positive (+) or negative (-) antigenic selection for $p < 0.05$ and strong antigenic selection (++, –) for $p < 0.01$, respectively.

3.3 Statistical and optimization methods

3.3.1 ROC curves

The Receiver Operating Characteristic (ROC) - curve [13] is a method of assessing and optimizing analysis strategies. The ROC curve shows the dependence of efficiency on the error rate.

For a given cut-off value which separates data points into a binary classifier system, the two variables sensitivity (true positives / total positives = true positive rate (TPR)) and 1-specificity (false positives / total negatives = false positive rate (FPR)) are calculated. By systematically varying the cut-off value from the lowest to the highest in the available range, a ROC curve can be plotted. Since TPR is equivalent to sensitivity and FPR to 1 - specificity, the ROC graph is sometimes called the *sensitivity vs 1 - specificity plot*. Prediction quality is measured by the area under the curve (AUC), which would be 0.5 for random predictions and 1.0 for perfect predictions.

3.3.2 Transinformation

Transinformation or mutual information is a part of information theory which gives the strength of the statistical connection (mutual dependence) of two random variables. Formally, the mutual information of two discrete random variables X and Y can be defined as:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{3.11}$$

Here $H(X)$ and $H(Y)$ denote the marginal entropies and $H(X,Y)$ is the joint entropy. In information theory, the marginal entropy is a measure of the uncertainty associated with a random variable. The marginal entropy is defined as follows:

$$H(X) = -\sum_{x \in X} p(x) log(p(x)) \tag{3.12}$$

where p is the probability that the event x occurs.

For the joint entropy each pair of possible outcomes (x, y) is considered. If each pair of outcomes occurs with probability $p(x, y)$, the joint entropy is then defined as:

$$H(X,Y) = -\sum_{x \in X, y \in Y} p(x,y) log(p(x,y)) \tag{3.13}$$

The most common unit of measurement of mutual information is the bit; this means that logarithms of base 2 are used.

The value for the transinformation I lies between zero and infinity. In order to rank the value of the transinformation, it is normalized to the contingency coefficient R, which is defined between zero and one (Equation (3.14)):

$$R(X,Y) = \frac{\sqrt{1 - \exp(-2 * I(X,Y))}}{\sqrt{1 - \exp(-2 * \min(H(X), H(Y)))}} \tag{3.14}$$

Note that, if the random variables X and Y are independent, then $H(X)+$ $H(Y) = H(X,Y)$ and therefore $R(X,Y) = 0$. The larger the value of the contingency coefficient R, the more information the two variables have in common. The contingency coefficient is 1 if a random variable can be completely calculated from the other.

3.3.3 Resampling - Bootstrap

Bootstrapping is a method of resampling in statistics. Intended parameters are calculated repeatedly on the basis of a single set of samples, as the underlying theoretical distribution may be unknown. It is a practice of estimating properties of a parameter, e.g. the variance of a sample set.

The method for the estimation of variance can be described as follows:

1. Sample $n-1$ observations with replacement randomly.
2. Estimate the mean value of the sample.
3. Repeat this bootstrap sampling for a large number (e.g. 2000 times).
4. Estimate the variance of the underlying dataset by using the estimated mean values.

3.3.4 Simulated annealing

Simulated annealing (SA) is a heuristic optimization procedure [67]. The procedure is used for detection of an approximate solution in optimization problems, when direct mathematical procedures and exhaustively trying out of all possibilities are not possible due to their high complexity.

The basic idea comes from the simulation of a cooling process. On heating up a metal the slow cooling provides enough time for the molecules to form stable crystals. Thereby a low-energy state could be reached, nearby the optimum.

Analogously to this physical process, each step of the SA algorithm replaces the current solution by a random solution in its neighborhood, chosen with a probability depending on the difference between the corresponding solutions and on a global parameter T (called the temperature), which is stepwise decreased during the process. For high temperatures (large T) the current solution changes almost randomly, but steadily towards the global optimum as T reduces to zero. The allowance for disadvantageous solutions prevents the method from being trapped at local optima.

3.4 Source code and datasets

The complete datasets and all tools including source code of all investigations presented here can be obtained by the author.

- Pool, a Java Applet containing nearly all algorithms presented here
- Lisa, a Java Application for the design of peptide libraries
- all Genespotter result files
- all AFFI tables
- all collected and aligned $VH_{186.2}$ sequences

Chapter 4

Results and Discussion

4.1 Reliability of array-based measurement of peptide binding affinity

4.1.1 Spot signal intensities: reproducibility and improvements

Performing a binding experiment with a cellulose membrane-bound peptide array (see Materials and Methods) results in measurable spot signal intensities (SI) signifying peptide/protein interactions (Figure 4.1). Standard deviation of the SIs measured for several peptide replicas on a membrane varies in the range from 8% to 22%. It is assumed that inhomogeneity of the modified cellulose membrane is largely responsible for this variation, since the synthesis quality of all peptides used in the binding experiment is well established [55, 72, 73, 109, 135].

The variations are quite reasonable compared to other high-throughput methods, but improvements on getting the inhomogeneity under control were to be developed.

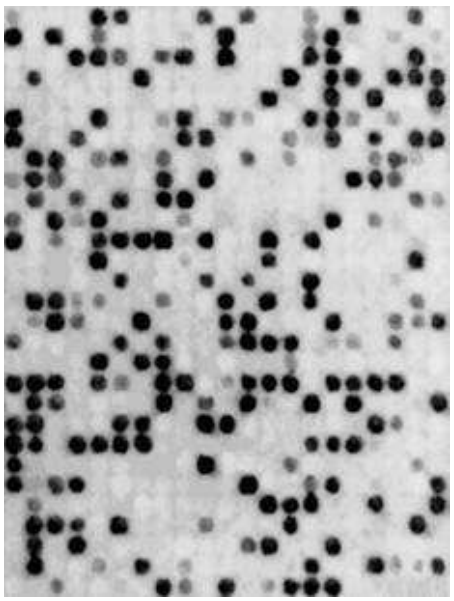


Figure 4.1: The membrane *SCM*-10. The membrane SCM-10 (8 cm x 8 cm) harbors 607 generated spots representing 38 peptides incubated with CB4-1 antibody.

4.1.2 Softimprovement

The membrane-related inhomogeneity of measured SIs was confirmed by analyzing their spatial distribution (Figure 4.2). For this purpose, the mean values of the replica SIs for each peptide on the membrane $CM_{6000} - 50$ were determined and the relation of each single SI to its mean value was calculated, thus determining regional trends. This revealed that peptides with similar mean signal intensities tend to have similar regional trends and peptides with different mean signal intensities have different regional trends (Figures 4.2B, D). Here, regional trends denote the trends of spots in differ-

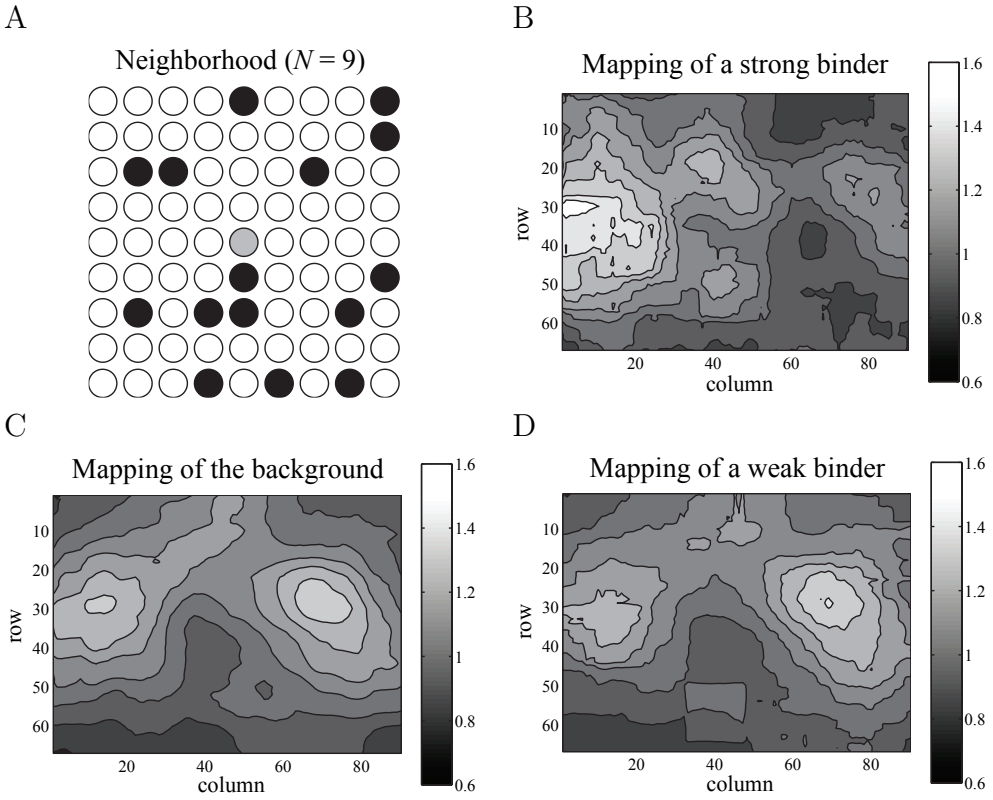


Figure 4.2: Neighborhood of spots. (A) Neighborhood of a spot (gray). The size of the neighborhood N indicates that it consists of N x N spots. The black spots represent reference spots. (B)-(D) Mapping of the membrane $CM_{6000} - 50$ (18 cm x 24.4 cm) containing 6000 spots (67 rows x 90 columns). Light areas denote relatively high, dark areas denote low values for the signal intensities. (B) Relative signals of the high affinity reference peptide (about 1000 data points, $pK_{dis} = 9$), (C) the background (6000 data points), (D) a weak binding peptide (about 1000 data points, $pK_{dis} = 4$). Note that maxima and minima of mapping (B) have different localizations compared to (C) and (D). The scale gives the relation between the local and the global mean signal.

ent local parts of the membrane. Furthermore, signal intensity fluctuations of the background correlate with those of weak binders (Figures 4.2C, D). After analyzing several membranes it became obvious that regional trends are intrinsic characteristics of individual membranes. This presents an opportunity to correct measured SIs and thus reduce standard deviation. A minimum of two references are necessary - one peptide with high affinity, and one with low affinity to the binding partner. The algorithm for improving measured SIs involves the following five steps:

1. calculating the mean SI values of all high ($\overline{H}_{global}$) and then low affinity references ($\overline{L}_{global}$).

2. defining a neighborhood i for each spot with N spots around it (Figure 4.2A).

3. calculating the mean SI values of the high- ($\overline{H}_{local,i}$) and low affinity references ($\overline{L}_{local,i}$) within each neighborhood.

4. calculating the variables a and b with the equations $\overline{L}_{global} = a_i * \overline{L}_{local,i} + b_i$ and $\overline{H}_{global} = a_i * \overline{H}_{local,i} + b_i$.

5. the improved signal intensity ($SI^{improved}$) for each spot in its neighborhood i can then be calculated from the measured signal intensities ($SI^{measured}$).

$$SI^{improved} = a_i * SI^{measured} + b_i \tag{4.1}$$

The standard deviation of the $SI^{improved}$ values for replicas of the same peptide will be smaller than that of $SI^{measured}$. As an alternative to a low affinity reference peptide, the background signals of each spot can also be used as references. The resulting standard deviations are shown for the SCM-50 membrane in Figure 4.3A (stars). The efficiency of improvement is influenced by the neighborhood size N, and the percentage of reference peptides regularly distributed on the membrane. To determine optimal parameters, standard deviations of $SI^{improved}$ were calculated for the CM_{6000}-50 membrane using background signals as the low affinity reference and considering different neighborhood sizes N and different percentages of high affinity reference peptide replicas (Figure 4.3B). The peptide GATPEDLNQKLAGN (Table 3.1) serves as high affinity reference. As expected, increasing the number of high affinity reference peptides improves the efficiency of the method and reduces the optimal neighborhood size N (Figure 4.3B). The results suggest that good adjustment for regional trends can be achieved when around

4% of the spots on a membrane are high affinity reference peptides, and by using a neighborhood with N x N spots, when N is between 10 to 20.

In practice, there are often no replicas available on a membrane, or no high affinity peptide is known beforehand. In such cases, $SI^{improved}$ can be calculated by using only the background signal around each spot according to

$$SI^{improved} = SI^{measured} * \frac{\overline{L}_{global,i}}{\overline{L}_{local,i}} \tag{4.2}$$

The improvements achieved by applying Equations (4.1) and (4.2) are compared in Figure 4.3C. Note that the decrease in standard deviation is smaller when using Equation (4.2), especially for peptides with high signal intensities.

SI measurements made from two identical membranes prepared with the same peptides, the same number of spots, the same concentration of antibody, the same FQ and the same incubation time were compared. Signal intensity deviations of a factor around 10 were found. Nevertheless, the relations of the signals were fairly constant for all peptides suggesting that, in order to compare SIs from different membranes, these values should first be normalized using

$$SI^{norm} = \frac{SI^{measured} - SI^{min}}{SI^{max} - SI^{min}} \tag{4.3}$$

Note that below, this equation will only be used for better visualization of the intermediate region (Figure 4.4), with SI being the mean signal intensity values for each recurrent peptide.

4.1.3 Correlation between signal intensities and dissociation constants

A representative repertoire of 38 peptides was selected from a set of over 100 peptides, all known to interact with mAb CB4-1 (Table 3.1). All chosen peptides are well described in the literature and dissociation constants (pK_{dis}) for the various peptide/antibody complexes span a range of more than five orders of magnitude ($pK_{dis} = 9.0$ to $pK_{dis} = 3.5$). As an example, mean values of measured SIs obtained from a SCM-10 membrane were used to draw an SI/pK_{dis} scatter plot (Figure 4.4A). In order to describe the SI/pK_{dis} dependency mathematically, as well as understand which parameters influence the correlation, a mass action law model was defined. It is based on the assumption that the system is in equilibrium and that the mass action law is

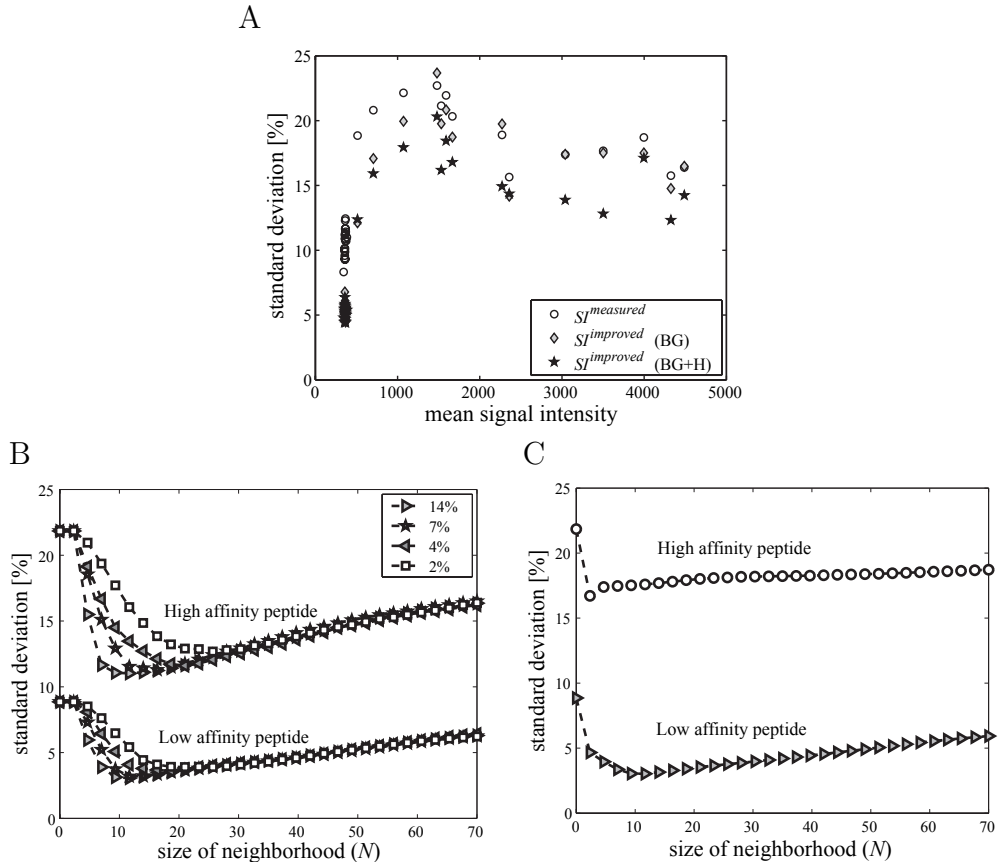


Figure 4.3: Standard deviation of the signal intensity and application of the noise reduction algorithm. (A) Standard deviations of the signal intensities for each of the 38 different peptides (10-15 spots each) on the membrane SCM-10 vs. their mean values (circles). The gray diamonds denote the standard deviation after application of the noise reduction algorithm using only background SI values as reference. The black stars show the effect of the noise reduction algorithm using background SI values as well as high affinity peptide repeats (14% of total spots) as reference. The obtained standard deviation of the corrected signal intensities depends on the size of the chosen local neighborhood (N) as well as on the number of reference peptides in it. Figures (B) and (C) show the standard deviation of corrected signal intensities for two peptides, one with low and one with high binding affinity, using the background as well as a high affinity reference peptide (B), respectively only the background as reference (C). In (B), four cases are shown where the number of high affinity reference peptides is 14%, 7%, 4% and 2% of the total number of spots on the membrane. The signals stem from membrane CM_{6000}-50.

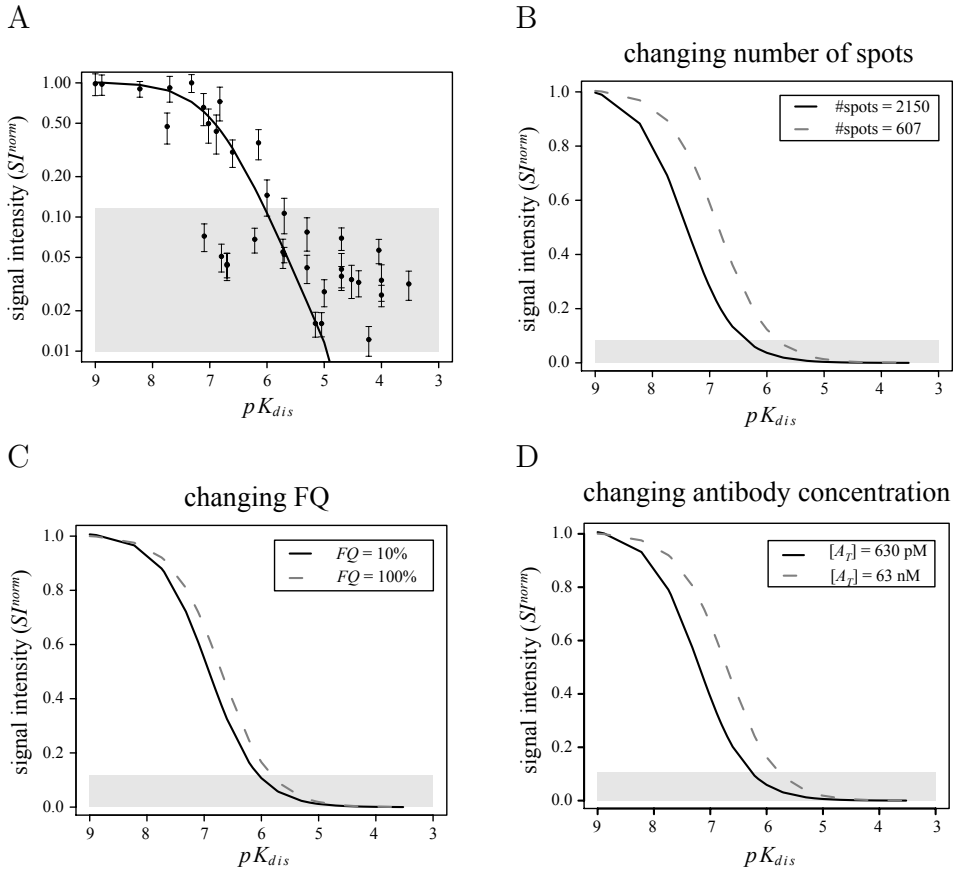


Figure 4.4: Mean signal intensities vs. dissociation constants of all measured peptides. All signal intensities are normalized according to Equation (4.3). The gray region marks all signals that are weaker than the mean background signal. (A) Mean values of the 38 different peptides on the SCM-10 membrane with an antibody concentration of 6.3 nM. The solid curve stems from fitting of Equations (4.4 - 4.7) to data. The scale of the SI-axis is logarithmic. (B)-(D) Fittings of the results of several experimental setups with all except one parameter held constant: (B) Two membranes prepared with a different amount of replicas of the peptides, (C) two membranes loaded with different FQ, (D) two membranes incubated with different antibody concentration.

the underlying effect. Furthermore, one has to take into account that competition among the peptides for mAb CB4-1 might take place, especially in cases when peptide concentrations or the number of spots are high compared to the antibody concentration.

Since cellulose membranes are porous, the concentration of antibody within the fibers may be different to the antibody concentration $[A_T]$ in the bulk solution. The inter-fiber spaces could be in the order of magnitude of antibody molecules in size. As described by Minton [92], this could lead to confinement effects and modified diffusion lengths. Since peptide/antibody interactions primarily occur in the inter-fiber spaces, a total effective antibody concentration $[A_T^{eff}]$ for this space was defined. Furthermore, the surface density of accessible peptides on a spot is not known. It may be influenced by positive cooperativeness: cluster formation of several peptide molecules within a spot may decrease its concentration [91, 93]. A total effective peptide concentration per spot $[P_T^{eff}]$ was defined, and assumed that it is the same for all spots. Unfortunately, it is not possible to determine the parameters $[P_T^{eff}]$ and $[A_T^{eff}]$ experimentally. Therefore these effective parameters were estimated by the proposed mass action law model. To formulate the model, Equations (4.4 - 4.6) were defined first. Here, $[P_i^{eff}]$ and $[A^{eff}]$ denote the concentrations of unbound peptide within a spot i $(i = 1, ..., n)$ and unbound antibody, respectively.

$$[A^{eff}] * [P_i^{eff}] = K_{dis_i} * [AP_i] \tag{4.4}$$

$$[P_T^{eff}] = [P_i^{eff}] + [AP_i] \tag{4.5}$$

$$[A_T^{eff}] = [A^{eff}] + \sum_{i=1}^{n} [AP_i] \tag{4.6}$$

Equation (4.4) simply states the mass action law for one single spot, Equation (4.5) sums the unbound and bound fractions of the peptide population in spot i and Equation (4.6) sums the concentration of free antibody with the sum of all complexes to the given total antibody concentration. This describes a nonlinear system of $2n + 1$ equations with the $2n + 1$ variables $[P_i^{eff}]$, $[AP_i]$ and $[A^{eff}]$ and the parameters $[A_T^{eff}]$ and $[P_T^{eff}]$. Furthermore, it is reasonable to assume that the concentration of the antibody-peptide-complex $[AP_i]$ within a spot is proportional to its measured signal intensity and to introduce the background signal as an additive factor. Thus, Equation (4.7) with the proportionality factor a and the background parameter b was defined additionally:

No.	Membrane	Spots	FQ	$[A_T]$	$\frac{a}{10^{12}}$	b	$[P_T^{eff}]$	$[A_T^{eff}]$	Corr
1	SCM-10	607	10%	6.3	52	2121	280	180	0.88
2	SCM-25	607	25%	6.3	663	3106	136	187	0.86
3	SCM-50	607	50%	0.63	87	1444	125	73	0.89
4	SCM-50	607	50%	6.3	74	2764	89	189	0.86
5	SCM-50	607	50%	63	69	2532	95	261	0.85
6	SCM-100	607	100%	6.3	100	2390	42	189	0.83
7	CM_{2090}-50	2090	50%	6.3	10	295	104	104	0.91

Table 4.1: Several fitted and experimental parameters for different experimental setups. The unit of the antibody concentrations ($[A_T]$, $[A_T^{eff}]$) is always nM, the peptide concentration $[P_T^{eff}]$ is given in pM.

$$SI_i = f\left(pK_{dis_i}\right) = a * [AP_i] + b \tag{4.7}$$

Equations (4.4 - 4.7) are used to fit the free parameters $[A_T^{eff}]$, $[P_T^{eff}]$ and b to the measured experimental data using the functions "nls" and "nls.lm" from the R language environment [102]. Parameter a was excluded from the fitting procedure in order to avoid redundancies and due to the fact that a differs from membrane to membrane. Instead, it was estimated by the formula $a = \frac{SI^{max}}{[P_T^0]}$, where $[P_T^0]$ denotes the sweeping start parameter of the fitting process for $[P_T^{eff}]$.

To evaluate the relevance of the model, several experimental parameters were varied specifically: the number of spots on a membrane, the FQ, and the antibody concentration. The fitting process was accomplished for each binding experiment. The correlation coefficients between the mean values of the measured SI and the fitted SI were found to be between 0.83 and 0.91. Table 4.1 summarizes all fitting results and the experimental conditions used. Figures 4.4B - 4.4D show the best-fits to the obtained experimental data after normalization with Equation (4.3).

The first observation is that the effective peptide concentration increases with decreasing FQ. Increasing the membrane FQ from 10% to 100% resulted in a decrease of the fitted total effective peptide concentration from 280 pM to 42 pM. This result is supported by the fact that absolute measured SIs decrease with increasing FQ as well. The decrease of effective accessible peptide concentration with increasing FQ can be explained by a clustering effect [91, 93], suggesting that much lower FQs than those used here could also yield good results.

Secondly, the fitted total effective antibody concentration $[A_T^{eff}]$ does not increase with the same order of magnitude as the concentration $[A_T]$ origi-

nally applied in the bulk solution. While $[A_T]$ increases by a factor of 100, $[A_T^{eff}]$ only increases by a factor of approximately four. Most likely, the total effective concentration of the antibody saturates within the membrane and one consequence is a nonlinear correlation between the total effective concentration within the spots and the concentration of the antibody in solution.

Thirdly, several competition effects were observed. Two membranes with identical antibody concentration and FQ and the same peptides, only with a different number of spots, i.e. different reiteration of the peptides (see Materials and Methods), show very different signal behavior. This can be understood if one considers that many spots with high affinity bind many antibody molecules, thus depleting the solution of free antibodies. Fewer antibody molecules are then available for binding to peptides with lower affinity, so they will have lower SIs compared to when there are only a few high affinity spots and therefore less competition for antibody (Figure 4.4B).

Furthermore, competition for antibodies may also possibly explain an effect described by Kramer *et al.* [73]: signal intensities either increase or decrease with the FQ of the membrane, depending on the peptide investigated. In a simulation of this experiment applying Equations (4.4 - 4.7), two model results were compared where only the peptide concentration was varied: $[P_T^{eff}] = 100\,nM$ and $[P_T^{eff}] = 400\,nM$, both with $[A_T^{eff}] = 10\,\mu M$ and seven peptides ($pK_{dis} = 9.0$ up to $pK_{dis} = 3.0$) each replicated 15 times. The results shown in Figure 4.5 illustrate that peptides with high binding affinity show decreasing signal intensities for decreasing $[P_T^{eff}]$ (and hence increasing FQ), while the converse applies to peptides with low binding affinity.

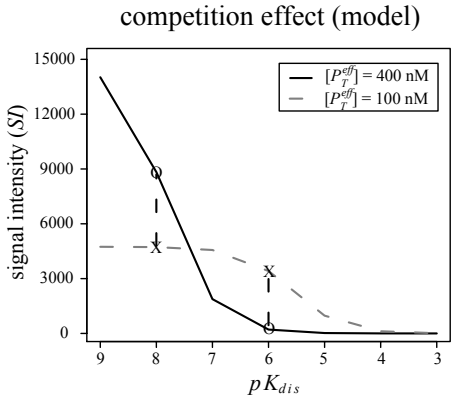


Figure 4.5: Influence of competition effects for resulting signal intensities. Signal intensity vs. dissociation constant calculated from Equations (4.4 - 4.7) for two cases which differ only in the peptide concentration. In this theoretical investigation the signal intensities are smaller for $[P_T^{eff}] = 400\,nM$ than for $[P_T^{eff}] = 100\,nM$ if peptides have low affinity, but larger if their binding affinity is high.

4.1.4 Detection of high-affinity binders from signal intensity data

The most important application of the SPOT synthesis technique is to detect peptides that have a strong binding affinity to defined targets. In practice, there are often no replicate peptides on cellulose membrane arrays so one can not calculate mean values. Therefore, the confidence of the SPOT synthesis technique in detecting high affinity binders for non-recurrent spots was investigated. First, classes of high and low affinity peptides had to be defined. As depicted in Figure 4.6A, two borders SI^1 and SI^2 on the y-axis (for the signal intensities) and two borders pK^1_{dis} and pK^2_{dis} on the x-axis (for the peptides' affinity) are defined. The four borders are free parameters defining three classes of signal intensities and three classes of dissociation constants, resulting in nine disjoint quadrants in the plot. For each of the nine quadrants the number of occurrences of spots can be assigned for each membrane and is denoted as n_{ij} ($i, j = 1, 2, 3$ are the enumerations of the SI classes and the pK_{dis} classes, respectively).

To determine how well the signal intensity class corresponds to the actual affinity class of the peptide, their mutual information is calculated (com-

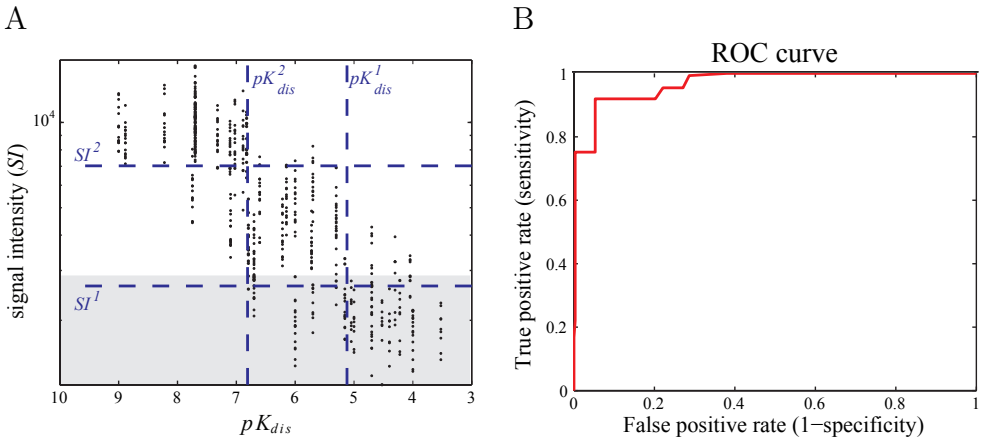


Figure 4.6: Classification of signal intensities into binding affinities. (A) Signal intensity vs. dissociation constant for each single spot. The class borders obtained by maximizing the contingency coefficient R (Equation (4.12)) for the classification in three affinity classes are shown as dashed lines. The gray region marks all signals that are weaker than the mean background signal. The SI-axis has logarithmic scale. (B) ROC curve generated for the classification of peptides of the membrane SCM-25 with $[A_T] = 6.3\,nM$ in two binding affinity classes. The SI-border was chosen as the mean value of the background signals plus three times its standard deviation (SNR=3), while the affinity border pK^1_{dis} was swept.

pare with Equations (3.11 - 3.14) in Materials and Methods). The entropies of signal intensity and dissociation constant (Equations (4.8 - 4.10)) and the transinformation of signal intensity and dissociation constant (Equation (4.11)) is estimated as follows:

$$H\left(SI\right) = -\sum_{i=1}^{3}\left(\frac{\sum_{j=1}^{3} n_{ij}}{\sum_{j,k=1}^{3} n_{jk}} * \log 2\left(\frac{\sum_{j=1}^{3} n_{ij}}{\sum_{j,k=1}^{3} n_{jk}}\right)\right) \tag{4.8}$$

$$H\left(pK_{dis}\right) = -\sum_{i=1}^{3}\left(\frac{\sum_{j=1}^{3} n_{ji}}{\sum_{j,k=1}^{3} n_{jk}} * \log 2\left(\frac{\sum_{j=1}^{3} n_{ji}}{\sum_{j,k=1}^{3} n_{jk}}\right)\right) \tag{4.9}$$

$$H\left(SI,\, pK_{dis}\right) = -\sum_{i=1}^{3}\sum_{j=1}^{3}\left(\frac{n_{ij}}{\sum_{k,l=1}^{3} n_{kl}} * \log 2\left(\frac{n_{ij}}{\sum_{k,l=1}^{3} n_{kl}}\right)\right) \tag{4.10}$$

$$I\left(SI,\, pK_{dis}\right) = H\left(SI\right) + H\left(pK_{dis}\right) - H\left(SI,\, pK_{dis}\right) \tag{4.11}$$

The contingency coefficient R is estimated by (Equation (4.12)):

$$R\left(SI,\, pK_{dis}\right) = \frac{\sqrt{1 - \exp\left(-2 * I\left(SI,\, pK_{dis}\right)\right)}}{\sqrt{1 - \exp\left(-2 * \min\left(H\left(SI\right),\, H\left(pK_{dis}\right)\right)\right)}} \tag{4.12}$$

The R values for a sweep of the parameters pK_{dis}^{1}, pK_{dis}^{2}, SI^{1} and SI^{2} are calculated and the borders with the best R value are determined.

This procedure was applied to several cellulose membranes and the resulting contingency coefficients lay between 0.83 and 0.96 (Table 4.2) showing high correlation for the distinction of the three defined classes of peptide binding affinities. SIs larger than SI^{2} may be associated with a dissociation constant smaller than pK_{dis}^{2} and SIs smaller than SI^{1} may be associated with a dissociation constant larger than pK_{dis}^{1}. The intermediate class between the two borders pK_{dis}^{1} and pK_{dis}^{2} isolates a small region of dissociation constants ("dynamic range"), so that measured signals inside the corresponding SI class may be associated with an approximate dissociation constant. Under the often used condition ($FQ = 50\%$ and $[A_T] = 6.3\,nM$), the dynamic range of the curve is between $pK_{dis}^{1} = 5.3$ and $pK_{dis}^{2} = 6.8$.

Applying the transinformation to more than the three classes described above led to singleton classes. This suggests that any other attempt to classify the results of the SPOT technology into more classes would be meaningless. For instance, it seems to be impossible to assign a dissociation constant to each measured signal. This is due to the extreme variability of the data in the central class and the saturation taking place far off the dynamic range.

No.	Membrane	$FQ(\%)$	$[A_T]^a$	R	SI^1	SI^2	pK^1_{dis}	pK^2_{dis}	AUC^b
1	SCM-10	10	6.3	0.92	4500	8900	6.8	7.2	0.97
2	SCM-25	25	6.3	0.92	3300	7400	5.3	6.8	0.97
3	SCM-50	50	0.63	0.96	3700	9200	6.8	7.2	0.98
4	SCM-50	50	6.3	0.90	3700	6500	5.3	6.8	0.96
5	SCM-50	50	63	0.93	3200	7600	5.3	6.8	0.95
6	SCM-100	100	6.3	0.83	3200	4800	5.3	6.8	0.90

[a] The concentration of $[\boldsymbol{A_T}]$is given in nM.
[b] For calculating the AUC value the SI-border of the ROC curve was the mean value of the background signals plus three times its standard deviation (SNR=3, see Figure 4.6B)

Table 4.2: Calculated pK_{dis} and SI-borders with their corresponding contingency coefficients (R) and AUC values.

In the data investigated the lower border for signal intensities (SI^1) never exceeds the background signal plus three-times its standard deviation, which is equivalent to a signal-to-noise ratio (SNR) of three. In order to confirm that this value is an objective border for distinguishing between the two classes of high and low affinity peptides ROC-statistic to the measured data was applied. Contrary to above, only one border pK^1_{dis} as the border distinguishing between low and high binding affinity classes need to be defined. Furthermore, the border for separating low and high signal intensities (SI^1) is determined with the criteria described above (SNR=3), finally resulting in four quadrants n_{ij} $(i, j = 1, 2)$. With these, true positive (TP) rates and false positive (FP) rates ($TP = \frac{n_{11}}{n_{11}+n_{21}}$, $FP = \frac{n_{12}}{n_{12}+n_{22}}$) are calculated. Moving the border pK^1_{dis} over the range of all possible values will change TP and FP. Plotting (TP, FP) - points into a diagram and connecting those gives a ROC curve (Figure 4.6B). Choosing the SI^1-border with SNR=3 results in AUC values of up to 0.98 for the differently synthesized membranes (Table 4.2), suggesting that signal intensity is an excellent classifier for high and low affinity binders. Note that the best pK^1_{dis} value for the distinction of high and low affinity is defined by the experimental setup and will usually not be known by the experimenter. The example ROC curve for membrane SCM-25 shown in Figure 4.6B suggests a good pK^1_{dis} value at 6.7: 90% of peptides classified as having high affinity in fact belong to the class of peptides with $pK_{dis} < pK^1_{dis}$, while only 7% of the peptides in this class were wrongly classified as being low affinity binders ($pK_{dis} > pK^1_{dis}$). An even higher accuracy of classification can be achieved when mean SI values of replica peptides are used.

4.2 Establishment of a substitution matrix based on binding affinity only

4.2.1 Data basis

Liying Dong analyzed in her PhD thesis 68 monoclonal IgG antibodies with respect to binding affinity to their corresponding epitopes and to epitope-homologous peptides with single amino acid (AA) substitutions [30]. For this purpose she applied substitution analysis using SPOT synthesis (Figure 4.7) as described in Materials and Methods (Section 3.1.4). As pointed out in the previous chapter the outcome of this technique underlies large variance, which however can be handled as described, enabling to distinguish with high significance between successful binding and loss of binding between peptide and antibody. It was shown that a dissociation constant of approximately 10^{-6} M marks the threshold between binding and non-binding. In a substitution analysis (Figure 4.7) each amino acid position of a peptide is substituted by every single one of the 20 studied amino acids leading to a complete profiling of antibody binding behavior. Wildtype epitopes always bind with high affinity to the antibody and therefore this method provides a binary measure for binding loss or conservation.

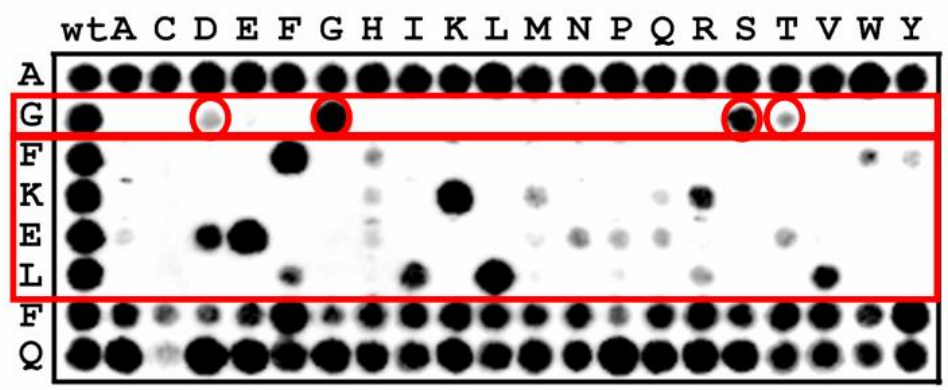


Figure 4.7: Substitution analysis of the peptide AGFKELFQ against the monoclonal antibody MiB5. Each spot represents a peptide with maximally one amino acid substitution compared to AGFKELFQ. A row represents all available substitutions of the respective amino acid. In the first row alanine is substituted, in the second row glycine and so on. The first column denotes the wildtype - the original peptide - the other columns denote the respective substitution. Dark spots indicate strong binding affinity. Here alanine is exchangeable without loss of binding affinity, while glycine is merely substitutable against aspartic acid, serine, threonine and itself, therefore it has a flexibility of four.

Here, definitions are given as follows:

1. Substitutions leaving the binding behavior of the epitope unchanged are called *conserving* substitutions.

2. Substitutions leading to a loss of binding are called *harmful* substitutions.

3. Furthermore, a residue within the epitope is called a *key residue* of *flexibility* n (KR^n), if it has n conserving substitutions (compare with Figure 4.7: $n = 4$ "black spots" in a row denote a key residue of *flexibility* four.).

Distribution of the amino acids in the dataset

The distributions of amino acids and key residues in the underlying dataset (Table 4.3 and Figure 4.8) give hints regarding the importance of amino acids in epitopes. Cysteine and methionine occur relative rarely in epitopes and also in key residues, while aspartic acid, proline and arginine seem to be very important in both. The amino acids alanine and phenylalanine show opposite properties. Alanine appears more often than expected in epitopes for uniformly distributed amino acids, but not very often as a key residue (38% of all occurrences). The opposite is the case for phenylalanine: if it is part of an epitope then it is in 77% of the cases a key residue.

AA	**# Occurrences**	**# $\mathrm{KR}^{\leq 15}$**
A	34.0	13.0
C	7.0	5.0
D	47.0	28.0
E	37.0	16.0
F	26.0	20.0
G	39.0	20.0
H	15.0	8.0
I	20.0	11.0
K	43.0	18.0
L	33.0	18.0
M	10.0	3.0
N	30.0	14.0
P	47.0	25.0
Q	30.0	10.0
R	38.0	24.0
S	29.0	12.0
T	42.0	16.0
V	35.0	16.0
W	16.0	11.0
Y	19.0	15.0
$\sum$	597	303

Table 4.3: Distribution of the frequency of all amino acids within the 68 epitopes and the frequency of those amino acids that are estimated as key residues with flexibility 15 or less.

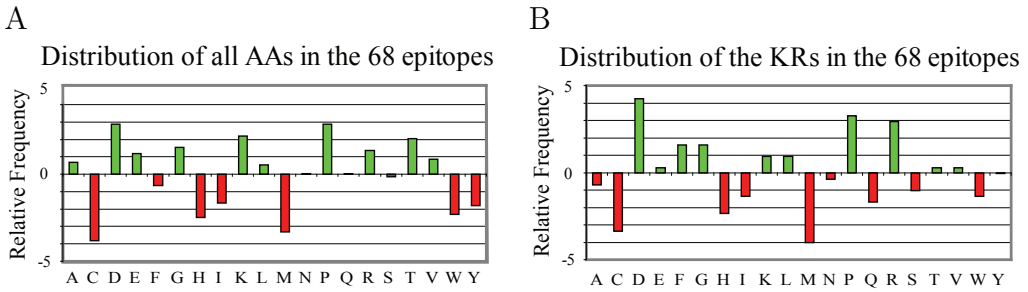


Figure 4.8: Distribution of amino acids and key residues in epitopes. The relative frequency of amino acids (A) and key residues of flexibility 15 or less (B) in 68 peptides used by Dong for substitution analysis are shown. The zero line on the y-axis denotes the expected occurrence of uniformly distributed amino acids (5% each). The difference to the zero line is plotted.

4.2.2 Generation of the substitution matrix AFFI

Together, the dataset and the definitions made above allow for analysis of similarities between the 20 studied amino acids with respect to binding behavior.

For all amino acid substitution pairs the frequency with which an amino acid substitution of a key residue leads to a change in binding behavior is determined. Two 20 x 20 matrices are defined, where the rows and columns are represented by the 20 amino acids. The matrix entries contain the number of *conserving* substitutions (matrix C^n, Table A.1) and *harmful* substitutions (matrix H^n, Table A.2), respectively (see appendix A). The two matrices are symmetric as original epitope residue and mutants are not distinguished. For the estimation of the matrices the measurements of all key residues with flexibility n or less ($KR^{\leq n}$) are considered. Therefore these matrices also depend on this parameter which is denoted by the superscript. The matrix $AFFI^n$ (Table A.3), called the *AFFI*nity substitution matrix with flexibility n, contains the relative fraction of conserving substitutions (see Equation (4.13)). Each entry in the matrix describes the probability of two amino acids to conserve binding affinity in the case of substitution.

$$AFFI_{ij}^n = \frac{C_{ij}^n}{C_{ij}^n + H_{ij}^n} \tag{4.13}$$

The mean value of all entries of $AFFI^{15}$ including synonymous substitutions is 0.25, meaning that approximately each forth random substitution is a conserving one. If the synonymous substitutions are not considered the mean is 0.21. The conserving probabilities of AFFI decrease with decreasing flexibility. Random substitutions for $AFFI^{10}$ and $AFFI^5$ have the average

conserving probability of 0.16 and 0.10, respectively, without consideration of synonymous substitutions.

The theory of random networks [107] states that a certain critical fraction of neighbors of any functional molecule needs to be functional as well, in order to give a neutral network in sequence space, which in turn is assumed to be a precondition for evolution taking place [66]. The required fraction of neighbors is given by Equation (4.14) and depends on the size κ of the alphabet. Since $\kappa = 20$ for the studied amino acid alphabet one finds $\lambda_{crit}(20) = 0.146$.

$$\lambda_{crit}(\kappa) = 1 - \sqrt[\kappa-1]{\frac{1}{\kappa}}, \quad \kappa > 1 \tag{4.14}$$

The value for the fraction of neutral (=conserving) substitutions of the matrix AFFI^5 (with mean value 0.10, s.a.), is clearly below λ_{crit}, while AFFI^{10} (0.16) has comparable size. AFFI^{15} (0.21) clearly exceeds the critical value and suggests the existence of neutral networks of epitopes based on the underlying key residues covering large fractions of the hypercube (see also [116]). The neutral pathways ensure "connectivity" between the amino acids. For AFFI^5 the connectivity of amino acids is low in contrast to AFFI^{15}.

Evaluation of the reliability of AFFI

In this section, the prediction quality of AFFI depending on differently flexible key residues is compared with other common substitution matrices, namely Blosum62 and PAM250. Therefore, ROC curves are generated (see section 3.3.1), an advantageous measure not relying on a single arbitrarily chosen cut-off value for the prediction probability, that can be equally applied to all substitution matrices.

The AFFI matrices' ROC curves are generated with the associated underlying dataset of key residue substitutions in a cross-validation process. The dataset is partitioned into four subsamples, each containing 25% of the data. Of the four subsamples, a single subsample is retained as the test data for testing the model, and the remaining three subsamples are used as training data. AFFI is generated using the training dataset. The test data is used as input for the generation of the ROC curve (Figure 4.9) with the matrix entries of the generated AFFI as cut-off values. This cross-validation process is then repeated four times, with each of the subsamples used exactly once as the test data. The four results are averaged. The AUC values of up to 0.84 implies very good prediction quality for *conserving* substitutions. With increasing flexibility of the AFFI matrix the AUC values decrease. This is

mainly due to the fact that matrices based on high flexibilities have to predict binding conservation for key residues with diverse flexibility - which is even harder. The AUC values remain almost constant even for different randomly chosen subsamples. The ROC curves for the Blosum62 and PAM250 matrices were generated with the dataset based on the $KR^{\leq 15}$ residues. The AUC values of the AFFI matrices are the greatest followed by Blosum62 (0.61), PAM250 (0.57) and the Identity matrix (0.50), suggesting that AFFI qualifies well for binding affinity purposes (Figure 4.9).

Using bootstrapping with replacement implemented in the custom-made java tool "Pool" 2000 new datasets of key residues (with the desired flexibility) were generated, and based thereon the AFFI matrices were calculated. Then the standard deviations of the matrix entries could be calculated. Table A.4 gives an overview on the coefficient of variation in each matrix entry of $AFFI^{15}$. The maximum variation is 49% for the substitution between the amino acids aspartic acid and tryptophan. On average the standard deviation is 14%. The mean value and variance of all entries of $AFFI^{15}$ is then 0.25 ± 0.04. The average coefficient of variation remains almost constant

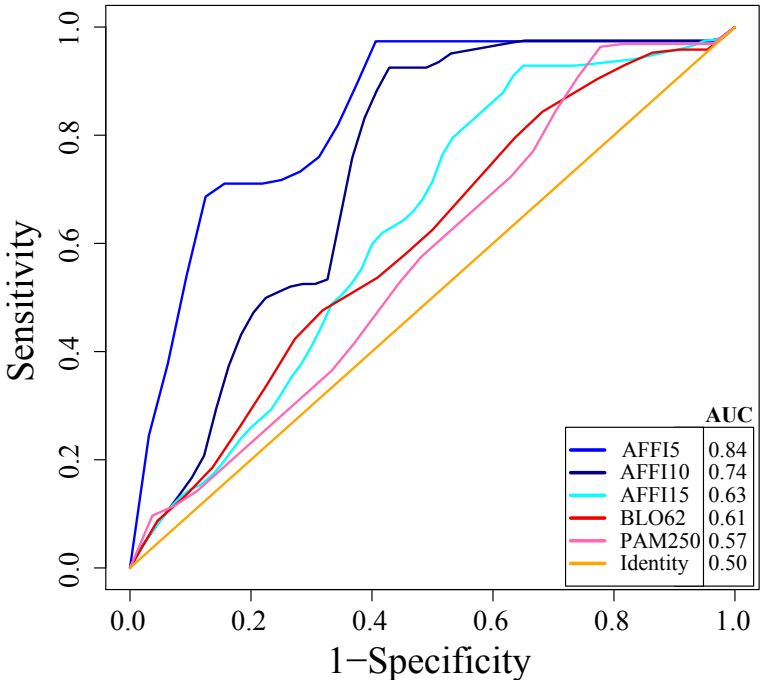


Figure 4.9: Reliability of AFFI. ROC curves of different substitution matrices. AFFI matrices were generated to cross-validate their prediction quality for binding conservation. The appropriate ROC curves of common substitution matrices are given for comparison. The AUC values of each ROC curve is given in the legend.

for different flexibilities, but the number of valid entries decreases with decreasing flexibility. The amount of available data is not sufficient to retrieve conserving substitutions for all amino acid pairs. AFFI^{15} for example has eight entries where no conserving substitution could be found. AFFI^{5} on the other hand has 118 entries without conserving substitutions due to the flexibility constraint. This was already indicated above by the low connectivity of AFFI^{5}.

Considering the underlying data basis, AFFI^{15} seems to be a good choice for further investigation of sequence space.

4.2.3 Grouping of amino acids

Conserving substitutions occur with higher probability between similar amino acids. In this section groupings of amino acids showing optimal arrangement are investigated.

For this purpose the 20 amino acids are randomly divided into k groups, which is called a partition. The number of ways of partitioning a set of n elements into k nonempty sets is described by the Stirling numbers of the second kind $S(n, k)$ [4]. For $n = 20$ and $k = 6$ the number of available different partitions is $S(20, 6) \approx 4.3 * 10^{12}$. An extension of the formula [15] considers additionally another constraint: the minimum number of elements l each group must contain. If for instance $l = 3$ the number of possible combinations would be $S(20, 6, 3) \approx 9.0 * 10^{10}$.

An objective function is necessary to evaluate the quality of a partition. The following two subsections introduce candidates for such a function, which will then be optimized and compared with each other. The parameters below are regularly used in the subsequent sections:

- n: number of used amino acids, usually $n = 20$ unless otherwise stated.
- k: number of groups within a partition.
- l: minimum number of amino acids within one group.

Clustering amino acids by proximity

The objective function presented in this subsection measures the average probability of conserving substitutions between amino acids appearing in the same group. As this function measures the relation between group members it is tagged *proximity measurement*.

Let C_i denote one group within a partition and p_{jm} the probability of a conserving substitution (matrix entries in AFFI) between amino acids j and

m. Firstly, all p_{jm} for all j, m belonging to one cluster within the partition are averaged according to Equation (4.15). Secondly, these results of all clusters are averaged. Equation (4.16) additionally takes the group sizes into account. This means that the conserving probability average of greater groups has more weight than the smaller ones.

$$p_1 = \frac{\sum_{i=1}^{k} \frac{\sum_{j=1}^{|C_i|} \sum_{m=1}^{|C_i|} p_{jm}}{|C_i|^2}}{k} \tag{4.15}$$

$$p_2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{|C_i|} \sum_{m=1}^{|C_i|} p_{jm}}{\sum_{i=1}^{k} |C_i|^2} \tag{4.16}$$

Since AFFI is a symmetric matrix $p_{jm} = p_{mj}$. All synonymous substitutions p_{jj} are equal to 1.

The calculation of $p_1[ILV]$ for the amino acid group $[ILV]$ is exemplified below:

$$p_1[ILV] = \frac{2 * (p_{IL} + p_{IV} + p_{LV}) + p_{II} + p_{LL} + p_{VV}}{9}$$

The value p_1 is the average of all summands. Conversely, p_2 first sums up all probabilities within every clusters of the partition and finally averages by the number of summands. If the partition would consist of groups having all the same size, then p_1 and p_2 would be identical.

In Figure 4.10 a random sample of one million partition implementations (realized with "Pool") with parameters $k = 6$ and $l = 3$ shows the distributions of p_1 and p_2 based on AFFI[15]. The analysis reveals gaussian distributed probabilities with mean values $\mu_{p_1} = 0.45$, $\mu_{p_2} = 0.44$ and identical standard deviations $\sigma_{p_1,p_2} = 0.02$.

Clustering amino acids by distance

The objective function analyzed in the following measures the average probability of conserving substitutions between amino acids appearing in ***different*** groups. This function measures the relation between amino acids in different groups, therefore it is tagged *distance measurement*.

p_3 averages the pairwise probabilities p_{jm} ***between*** the clusters, which are averaged again into the averaged distance probability of the defined partition. p_4 is the analogon to p_2, which also takes group size or better the number of interactions between the groups into account.

$$p_3 = 2 * \frac{\sum_{i=1}^{k} \sum_{j=i+1}^{k} \frac{\sum_{m=1}^{|C_i|} \sum_{n=1}^{|C_j|} p_{mn}}{|C_i| * |C_j|}}{k * (k-1)} \tag{4.17}$$

$$p_4 = \frac{\sum_{i=1}^{k} \sum_{j=i+1}^{k} \sum_{m=1}^{|C_i|} \sum_{n=1}^{|C_j|} p_{mn}}{\sum_{i=1}^{k} \sum_{j=i+1}^{k} |C_i| * |C_j|} \tag{4.18}$$

Generating a random sample of one million partition implementations yields a distribution of p_3 and p_4 with $k = 6$ and $l = 3$ (Figure 4.10). The analysis reveals gaussian distributions with mean values $\mu_{p_3,p_4} = 0.21$ and standard deviations $\sigma_{p_3,p_4} = 0.01$.

The difference of the mean values compared to p_1 and p_2 is mainly due to the fact that p_1 and p_2 do also consider synonymous substitution probabilities (which are equal to 1), whereas p_3 and p_4 do not. With decreasing k both the mean and variance of p_1 and p_2 decrease as well, approximating the respective values of p_3 and p_4. The effect of a decreasing l is an increase of the variances of p_1 and p_2 while the mean of p_1 increases and the mean of p_2 decreases. The mean of the distance measures (p_3 and p_4) remains constant while the variance slightly increases with decreasing l.

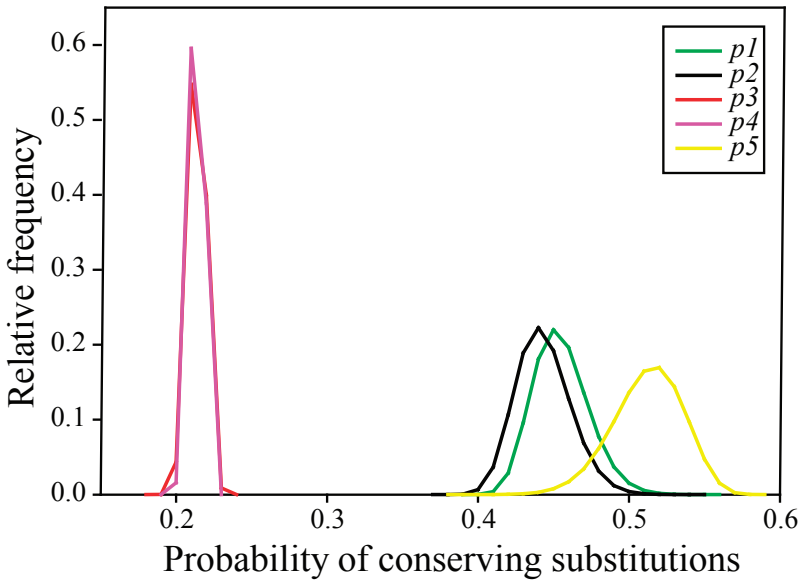


Figure 4.10: Probability distributions of conserving substitutions within amino acid partitions. The density functions of the average probability of conserving substitution is plotted for the distributions specified in Equations (4.15 - 4.19) with parameters $l = 3$ and $k = 6$.

Estimation of the optimal partitions

Discovering the optimal clusters, depending on the parameters k and l, is done by simulated annealing for each Equation (4.15 - 4.18) as objective function (implemented in "Pool"). For the Equations (4.15, 4.16) the maximum is wanted (=maximal proximity), while for Equations (4.17, 4.18) it is the minimum (=minimal proximity=maximal distance).

The best performing partitions according to the optimization of p_1, p_2, p_3 and p_4 are listed in the tables A.5 - A.12 (see appendix A) and illustrated for $k = 6$ and $l = 3$ in Figure 4.11.

The clustering based on p_1 shows an evident discrepancy compared to that based on p_2. While p_1 prefers many small groups and one big group, p_2 tends to create groups of the same size. The big group of p_1 may be called the "remainder" group as all the other groups are really optimized in showing big conserving properties among its members, while in the big

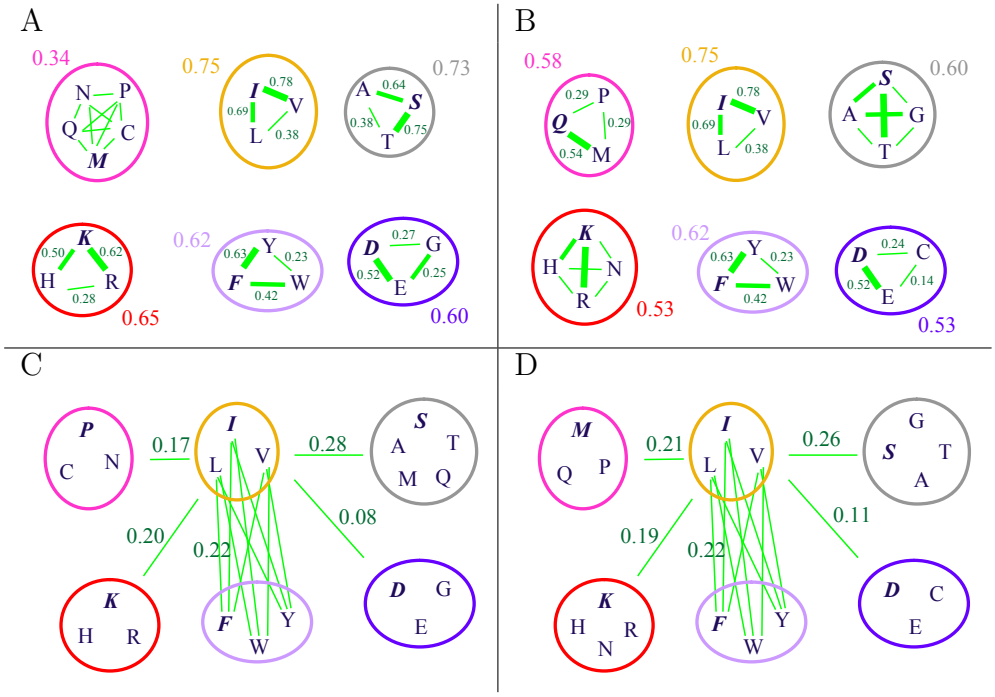


Figure 4.11: Groupings of amino acids for optimal proximity and distance partitions. Optimal arrangement of the amino acids obtained for the parameters $l = 3$ and $k = 6$ using the proximity of amino acids without (A, p_1) and with regard to group sizes (B, p_2) and the distance of amino acids without (C, p_3) and with regard to group sizes (D, p_4). The numbers in the figure indicate the probabilities for conserving substitutions (Equations (4.15 - 4.18), Tables A.5 - A.12). In figures (C) and (D) the average conserving probabilities to the other groups are illustrated for the [ILV] group only.

group amino acids are collected that have small conserving probabilities to all others. Thus, the amino acid space is split into two principal parts. One part consisting of the remainder group shows a rugged landscape and is therefore hard to cover, the other part indicates a very flat landscape, since its amino acids are exchangeable with very high probability. On the other hand p_2 divides sequence space into nearly equal sized parts (independent of parameter l) without a remainder group. There is a trend to equalize the average probabilities for conserving substitutions within the groups for p_2 compared to p_1 as illustrated in Figures 4.11A and 4.11B. The same is true for p_4 and p_3 (Figures 4.11C and 4.11D). Nevertheless, both of these optimal distance partitions produce remainder groups and there are also only slight differences between them. Interestingly, the proximity ($p_{3,4}$) and the distance partitions ($p_{1,2}$) divide the amino acid space similarly. The hydrophobic group [ILV] seems to be the most pronounced one, followed by the acidic [HKR] and the aliphatic [FWY] group. Furthermore, small polar alcohols [ST] and negatively charged amino acids [DE] tend to group together.

The tables with $l = 1$ (A.5, A.7, A.9, A.11) give hints for the amino acids that have the most harmful effect on substitutions. The amino acids W and C are the first that leave the "community". This is already apparent when looking at Table A.3: W and C are the amino acids with the most red colored cells indicating harmful substitutions.

The best performing partition according to the proximity clustering ($p_{1,2}$) is naturally obtained for $k = 20$ and $l = 1$ with $p_1 = p_2 = 1$. For distance optimization this partition is worst with $p_{3,4} = 0.21$, which is exactly the conserving probability for a random substitution within AFFI[15] when excluding synonymous substitutions. The results for p_2 show that the minimal group size (parameter l) does not play a role at all, besides two exceptions ($l = 1, k = 10$ and $l = 5, k = 4$) where l is a real constraint. This is due to the fact that the conserving probabilities are averaged by the group size. For the other distributions the parameter l is essential for the obtained results.

The best performing partition according to the distance clustering ($p_{3,4}$) is obtained for $k = 3$ and $l = 1$ with $p_3 = 0.09$. Here the overall most harmful substitutions W and C are excluded from the "community". For the proximity probabilities the resulting p value can be interpreted as the average probability of conserving substitutions, if substitution is only allowed within the groups, while the opposite is true for the distance probabilities $p_{3,4}$. Here p denotes the average probability of conserving substitutions, if substitution is only allowed between the groups.

Each resulting group can be interpreted as a main component of the sequence space. The higher the average probability of conserving substitutions within a group is the better the component represents the associated amino

acids. Vice versa, the lower the average probability between the groups is, the more distant are the groups within sequence space regarding binding properties of associated peptide sequences.

In the following section yet another objective function for partition optimization is introduced.

4.2.4 Reduction of the set of amino acids - selection of representative group members

After establishing of optimized groups of amino acids, each group can be regarded as a type of monomer, and can be represented by a single amino acid, yielding a reduced set of amino acids.

The most intuitive method for choosing a representative of a group is to take the amino acid that is "closest" to the other group members. This means that the chosen representative of a group has the greatest sum of probabilities of conserving substitutions to its group mates. If there are two or more members having the same maximum sum of probabilities (e.g. for groups with the size 2) the one is chosen having the greatest sum of probabilities compared to all other amino acids.

For example in the group $[ILV]$ the amino acid I has a conserving probability to V of $p = 0.78$ and to L of $p = 0.69$, while the conserving probability for L to V is $p = 0.38$. Therefore I is chosen as the representative of this group, called the "center of the group". The representatives of each group are highlighted in bold and underlined in the tables A.5 - A.14 in appendix A.

The optimal partitions of the distributions $p_1 - p_4$ are optimized according to overall group properties, but do not necessarily establish the optimal representatives partition. In order to find the optimum in that sense, Equation (4.19) is used as new objective function.

$$p_5 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{|C_i|} p_j^{R_i}}{20} \tag{4.19}$$

Here, $p_j^{R_i}$ denotes the conserving probability of group member j to its representative. A random sample of one million partition implementations with parameters $k = 6$ and $l = 3$ and the average probability p_5 based on AFFI[15] reveal a gaussian distribution with mean value $\mu_{p_5} = 0.51$ and standard deviation $\sigma_{p_5} = 0.02$, which is shown in Figure 4.10.

The best performing partitions according to the optimization of p_5 (simulated annealing with Equation (4.19) as the objective function to maximize)

are listed in the Tables A.13 - A.14 and illustrated in Figure 4.12. Here, the best performing partition again is obtained for $k = 20$ and $l = 1$ with $p_5 = 1$.

Similar to the results obtained for p_2, the minimal group size (parameter l) plays almost no role, it has only minor effects for relatively large l. The results depend only on parameter k and have a higher average conserving probability within the groups compared to $p_{1,2}$. Due to the facts listed above, the optimal partitions obtained by p_5 qualify for further practical exploitation. Before the performance of a practical application of these partitions can be investigated, multiple single-point amino acid substitutions have to be considered.

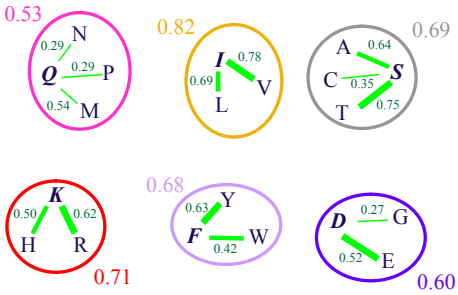


Figure 4.12: Groupings of amino acids for an optimal representatives partition. Optimal arrangement of the amino acids obtained for the parameters $l = 3$ and $k = 6$ using the representatives of the amino acids (p_5). The numbers in the figure indicate the probabilities for conserving substitutions.

4.2.5 Multiple single-point amino acid substitutions

So far only single-point amino acid substitutions were investigated. In this section multiple single-point amino acid substitutions and the resulting probability for conserving substitutions are examined.

Under the assumption that substitutions are independent, the probability of conserving binding behavior after multiple substitutions (p_{ms}) is simply determined by multiplying the single probabilities:

$$p_{ms}(p_1, ..., p_n) = \prod_{i=1}^{n} p_i \tag{4.20}$$

Starting from a wildtype epitope a random walk (excluding synonymous substitutions, $p = 0.21$ according to AFFI[15]) through the sequence space would give a probability for conserving substitutions of $p_{ms} < 1\%$ already after three random substitutions (Figure 4.13).

However, when considering an optimized reduced set of amino acids as developed in the previous section based on the representatives, meaning that

only substitutions within optimal partitions (e.g. representatives, p_5) are allowed, the average probability for a conserving substitution considerably increases. For $k = 6$ the average conserving probability is $p_5 = 0.67$ (Table A.13), and when neglecting synonymous substitutions it is $p_5 = 0.53$. In that case even after seven substitution steps the probability of conserving substitutions is $p_{ms} > 1\%$. That means that a complete substitution of a 7-mere along that optimal partition would have a greater conserving probability than random substitution at only three positions. This finding supports the idea for using a reduced set of amino acids with representatives as starting point for epitope search.

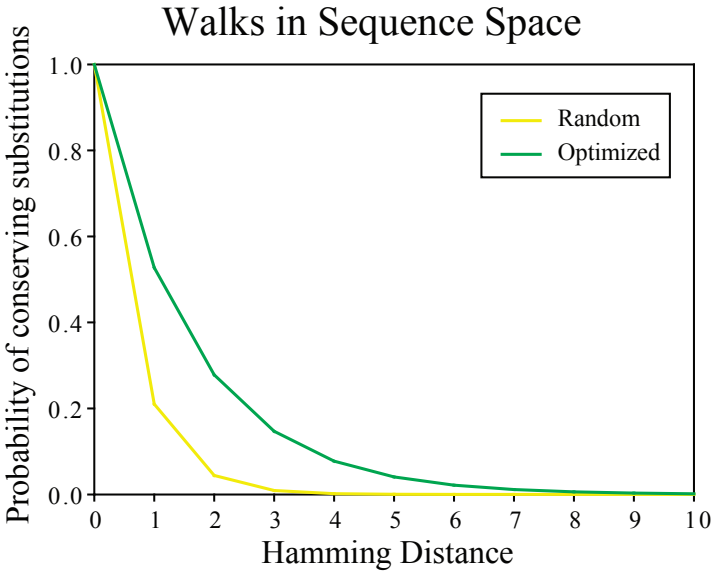


Figure 4.13: Multiple single-point amino acid substitutions. On the x-axis the number of substitution steps is given (Hamming distance), while the y-axis shows the expected probability of conserving substitutions for walks within the optimized partition groups based on representatives for $k = 6$ and random walks, respectively. For walks within the groups the probability of a conserving substitution is considerably higher.

4.2.6 In search of epitopes - performance of a reduced set of amino acids

For the experimental search of epitopes an optimal strategy is demanded with which minimal experimental effort is needed for the estimation of many different epitope candidates. Liying Dong [30] found that common linear epitopes usually consist of five to ten residues with two to five key residues. In order to completely cover the sequence space of a common 7-mer epitope,

one would need to synthesize 20^7 peptides. This is (at least at the moment) unachievable, even with modern high-throughput technologies. Using an optimized reduced set of seven amino acids also reduces the number of peptides to 6^7. Furthermore, it is observed that linear epitopes in general do not lose their binding affinity if they are embedded within a larger peptide. For example in 20-mer peptides 14 seven-mers are embedded. If 20-mer peptides are chosen carefully in the sense that all subsequences of length 7 do occur only once, the number of peptides needed to realize all combinations of 7-mers is reduced to $N = \frac{6^7}{14} \approx 20,000$ (illustration in Figure 4.14). Since very weak binding peptides can be identified using peptide arrays prepared by SPOT synthesis, it can be assumed that binding motifs with a reduced set of amino acids are detectable.

A disadvantage of the SPOT synthesis is that with longer peptides the error rate for falsely produced peptides increases as well. A second problem of long peptides is that they fold into secondary structures. An amino acid length of 20 is the maximal appropriate length (personal communication, Dr. Volkmer, AG Med. Immunologie, Charité Berlin).

Furthermore, it has to be clarified that according to the theory of random networks [107] there are in general no neutral pathways given for a reduced amino acid set. The required fraction of neighbors for a reduced set of e.g. $k = 6$ amino acids according to Equation (4.14) is $\lambda(6) = 0.301$. This is out of

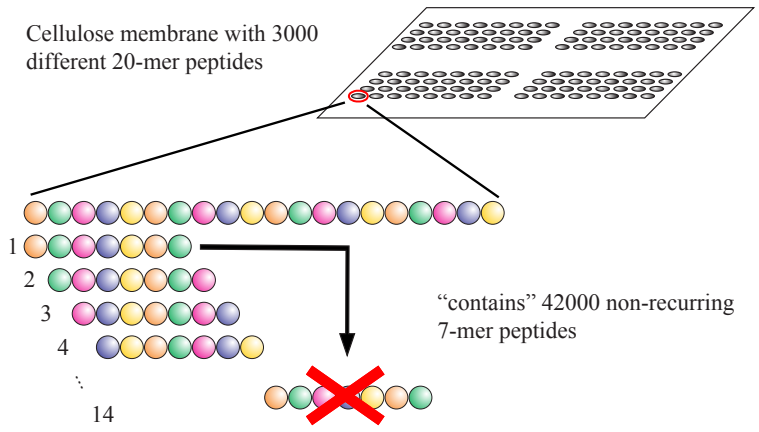


Figure 4.14: Optimized peptide library design. In a 20-mer peptide 14 seven-mers are embedded. Therefore, a library comprising 3000 20-mer peptides contains 42,000 subsequences of length 7. In the sequence space of 6 representatives containing $6^7 \approx 280,000$ peptides there is a high probability for 7-mers to occur more than only once therein when chosen randomly. In the optimized library design redundancy is avoided benefiting diversity.

reach - the theoretically expected fraction of conserving substitutions, which is the average distance between the different amino acid groups (distance optimization of partitions, section 4.2.3), is at most $p_4 = 0.18$ (Table A.12). The critical border of a reduced set is $k = 14$. Here, the required fraction of neighbors is $\lambda(14) = 0.184$, while the theoretically expected fraction of conservation is $p_4 = 0.19$ (Table A.11). Therefore, the affinity landscape of a sequence space of a reduced set of less than 14 representatives is too "rough" for evolution taking place on it. Evolutionary algorithms for estimation of binders will fail and the experimental detection of binding affinity for each single sequence will be indispensable.

Libraries based on a reduced set of amino acids vs. random libraries

One might ask why a library based on an optimized partition or a reduced set of amino acids respectively should represent the sequence space substantially better than for example a random library (Reineke et. al [109])? The question arises whether a library based on a reduced set of amino acids as described and provided above (p_5, Table A.13) has the same detection power or whether it needs more or less peptides for the detection of epitopes?

Here, definitions are given as follows:

- N: number of peptides within the library
- L: length of the synthesized peptide
- D: length of epitopes
- R: size of the used reduced set

The probability of having a recurrent epitope within a library that is optimized as illustrated in Figure 4.14 is then:

$$p_r(N, L, D, R) = \frac{N * (L - D + 1)}{R^D} \tag{4.21}$$

Reineke [109] used a random library with $N = 5525$ peptides each having the length $L = 15$. The epitope length may be defined as $D = 7$. That means that a search of all 7-mer epitopes is desired. Reineke's random library with all $R = 20$ amino acids has a recurrence probability of $p_r(5525, 15, 7, 6) = 0.00004$. Choosing a reduced set with $R = 6$ the probability increases to $p_r = 0.18$. If the conserving exchangeability of the representatives to their group mates would be 100%, meaning that a substitution does not alter

binding behaviour, 18% of the random library would be redundant. $N' = N * p_r = (1 - 0.18) * 5525 = 4531$ elaborate peptides would have the same detection power. But obviously the exchangeability is not 100% within the groups, except for the partition comprising 20 singletons.

The conserving probabilities of all amino acids to their representatives must be taken into account additionally. This probability p_{ms} is estimated via Equation (4.20) with the optimal partitions based on representatives (p_5). For $R = k = 6$ the representatives are [SEFKIQ] with an average conserving probability of $p_5(R = 6) = 0.67$ (Table A.13). The probability for a conserving substitution of the $R = 6$ letter alphabet within an epitope of length $D = 7$ is then $p_{ms}(D, R) = p(R)^D = 0.67^7 = 0.06$.

The total probability of having epitopes with identical binding behavior within the library computes as follows in Equation (4.22).

$$p_s(N, L, D, R) = p_r(N, L, D, R) * p_{ms}(D, R) \tag{4.22}$$

The result for the chosen parameters $N = 5525, L = 15, D = 7, R = 6$ is $p_s = 0.01$. Therefore it is expected to have 1% redundancy in the chosen library with the desired coverage of the $D = 7$-mer epitope space. This redundancy is dependent on several parameters.

What is the effect of the alteration of parameter *D*? For $D = 6$ the probability of having a recurrent epitope is $p_r(5525, 15, 6, 6) = 1.18$. So it is certain that there are identical epitopes within the library and the library size for that reduction scheme is oversized. But to be able to compare with other parameters, this has to be disregarded for a moment and the probability is cut to $p_r = 1$. With $p_{ms} = 0.67^6 = 0.09$, one gets $p_s = 1.0 * 0.09 = 0.09$, thus 9% of the chosen library is redundant for the 6-mer epitope space. With a decrease of the dimension of the desired epitope space the redundancy in the library increases.

What is the effect of the alteration of parameter *R* (keeping *D* = 6)? With decreasing R the number of needed peptides would rise again, because the probability of recurrence remains constant at $p_r = 1$ and the average probability of conserving substitutions within the partition decreases. For $R = 5$ and the representatives [SFKIQ] the probability for redundant information is $p_s = 1.0 * 0.63^6 = 0.06$. For increasing $R = 7$ and the representatives [SFKIQEA] the average conserving probability of the amino acids to their representatives is $p_5(R = 7) = 0.70$. The probability of having a recurrent epitope is $p_r(5525, 15, 6, 7) = 0.47$. With $p_{ms} = 0.70^6 = 0.12$ the result is then $p_s = 0.47 * 0.12 = 0.06$. Thus, the redundancy is still decreasing and furthermore there is an optimal number of representatives for each desired epitope space. For instance 5525 15-mers and the coverage

of the $D = 6$-mer sequence space, the best reduced set would be one with $R = 6$. Because of $p_r = 1.18$, the library size might additionally be reduced to $N' = 5525/1.18 = 4682$. Alternatively the redundant 843 peptides can be used to cover parts of the next dimension ($D = 7$). The same could and should be done with the additional 9% $\widehat{=}$ 421 redundant epitopes.

Summing up, the original random library contains avoidable redundancies or, more precisely, does not span sequence space in an equidistant way. Increasing the library size (N) or the length of the peptides (L), would increase the optimal R. For other epitope space dimensions (D), the optimal number of the reduced set and also the expected magnitude of the effect changes (Figure 4.15). For covering the epitope space of 4-mers 16 representatives may be used and nearly 80% of the original library can be economized yielding the same results. The larger the demanded epitope space the less representatives can be used and the less is the economization.

Generally speaking, depending on the available size of the library and the demanded epitope space, the optimal size of the needed reduced set should be fitted. If there will be the day, when 20^4, 20^5, 20^6, 20^7 or even 20^8 peptides can be technically handled and synthesized all at once, then all 20 naturally

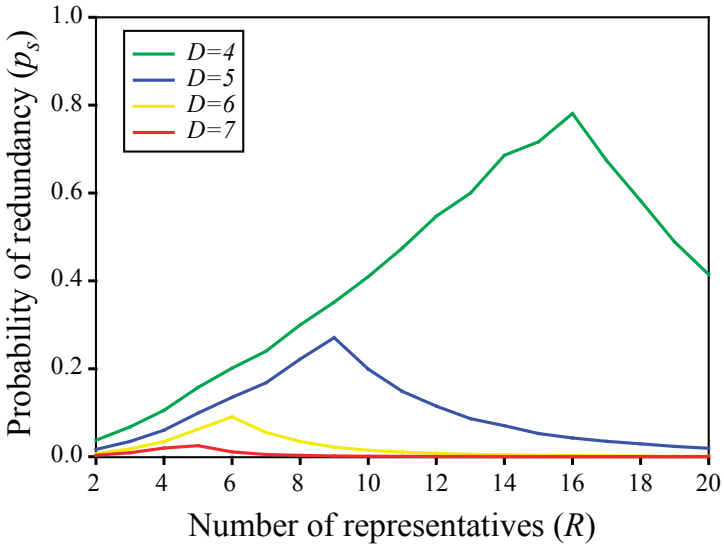


Figure 4.15: Probability of redundancy in random libraries (p_s) depending on the alphabet size (R) and the epitope length (D). The fixed parameters chosen are the library size $N = 5525$ and the peptide length $L = 15$. The redundancy of epitopes within a random library is dependent on the epitope length and the number of representatives. The larger the demanded epitope length the less representatives must be used in order to reach the maximal redundancy, which is decreasing with increasing epitope length.

occurring amino acids may be taken for the search of the sequence space - without any gap. Until then an optimized reduced set is always the better choice!

4.2.7 In search of epitopes - natural coverage of sequence space

In this section another application of the substitution matrix AFFI is presented. The basis are the DNA sequences that are the blueprint for antibodies within organizms. The immune system solves the problem of covering epitope space by using a diverse antibody repertoire. It is legitimate to ask how it does that. All antibody sequences originally stem from only a few genes: V(D)J and C genes. V genes provide a large and important part of the antibody sequence. They have the well defined regions CDR and FR. The CDR makes the contact to the antigen, which means that especially this region has to cover epitope space, while the FR takes care of structure and stability of the antibody and is therefore expected to be more conserved. The question is whether it is possible to visualize these features with AFFI?

For C57BL/6 mice the existing V genes are well known. The sequences can be obtained from the IMGT database [76]: 104 VH, 53 VL_κ sequences are available. As mice do only have few VL_λ chains, VL_λ chains are not considered in the following.

V gene repertoire

All n V gene sequences are aligned according to the IMGT sequence format and the conserving probability (CP) at each codon position i is pairwise computed according to the $AFFI^{15}$ matrix and finally averaged (CP_i) according to Equation (4.23) resulting in Figure 4.16. The term $AFFI^{15}(i,j,k)$ in Equation (4.23) stands for the conserving probability between the two amino acids located at the codon position i of the two V gene sequences j and k.

$$CP_i = \frac{\sum_{j=1}^{n}\sum_{k=j+1}^{n} AFFI^{15}(i,j,k)}{\frac{n(n-1)}{2}} \tag{4.23}$$

In the FR of the VH sequences are four static positions with no amino acid change, while there are none within the CDR. The average conserving probability in the CDR of the VH sequences is $CP_{CDR} = 0.60$ contrary to $CP_{FR} = 0.72$ in the FR. VL_κ sequences have eight completely conserved amino acids and they are again all within the FR. The average conserving probability of the VL_κ sequences is $CP_{CDR} = 0.60$ in the CDR in contrast

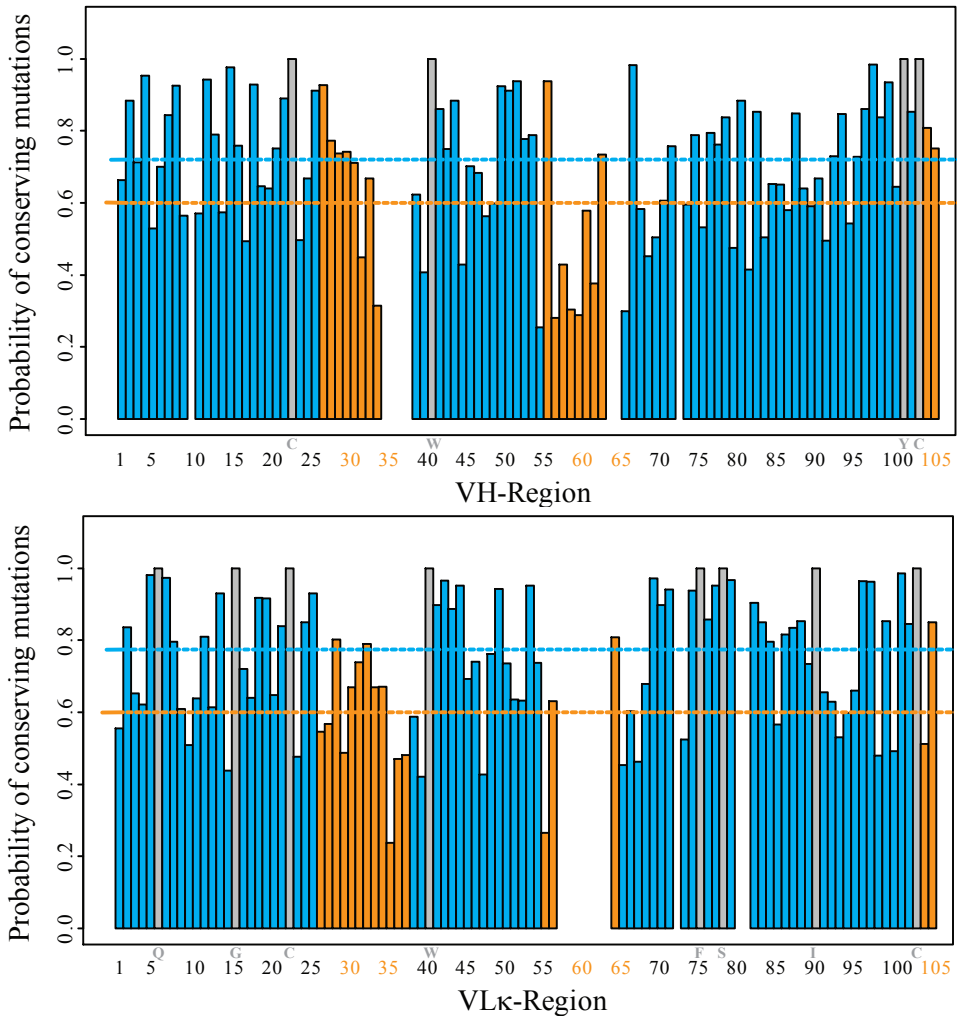


Figure 4.16: V gene repertoire of C57BL/6 mice - coverage of sequence space. V gene sequences of C57BL/6 mice are aligned according to the IMGT format. The probability for conserving substitutions is computed pairwise at each amino acid position according to AFFI[15]. Gray bars denote positions of total conservation, meaning that all sequences are identical at that position; the amino acid is indicated by the letter beneath. Orange bars are positions within the CDRs, cyan bars indicate positions in the FR. The average conserving probability in these regions is depicted by the dashed lines. The gaps in the sequence are due to the chosen IMGT format.

to $CP_{FR} = 0.77$ in the FR. The larger mean values in the FR compared to the CDR are in both cases significant according to the significance range of a one-tailed Mann-Whitney U test ($p = 0.03$ for the VH and $p = 0.001$ for the VL_κ chains).

The repertoire of V genes may be interpreted as a "coarse-tuning" step of the immune system to cover sequence space. Especially in the CDR harmful mutations are dominant indicating that sequence space is more widespread covered than in the FR. Despite the dominant harmful mutations in the CDR, the average conserving probabilities are still twice as common compared to a random mutation ($p = 0.25$). This indicates that there are still some position specific conservations in the CDR, especially in the CDR1 region of the VH chain. The CDR2 contains rather diverse amino acids.

Single-point somatic hypermutations

A further "fine-tuning" step is performed by the immune system after antigen contact via somatic hypermutation, during which codons are mutated usually once at most. Regarding the transition bias of V genes as described in Materials and Methods (see Figure 1.4), one can also quantify the outcome of this "fine-tuning" step in terms of conserving probabilities. The question here is, which region, FR or CDR, is preferred for the conservation of binding behavior?

The fine-tuning conserving probability of a point mutation within a codon (FP) is computed as the conserving probability of the two involved amino acids according to AFFI[15] weighted by the transition and the intrinsic mutability bias within the codon ($p_{m,k}$) as specified in Equation (3.6). The conserving probability for a STOP Codon is defined as zero. All possible single-point mutations of each nucleotide within a codon are considered and their results are averaged. Finally, the resulting fine-tuning conserving probability (FP_i) of a codon position is the average value of all available n V genes at this position according to Equation (4.24) resulting in Figure 4.17. The term $AFFI^{15}(i, j, k, l)$ in Equation (4.24) stands for the conserving probability between the amino acid of the V gene sequence j located at the codon position i and the amino acid resulting via a mutation of the k^{th} nucleotide into nucleotide l, where l is indexing the remaining three nucleotides.

$$FP_i = \sum_{j=1}^{n} \frac{\sum_{k=1}^{3} \sum_{l=1}^{3} p_{m,k} * AFFI^{15}(i, j, k, l)}{3n} \tag{4.24}$$

The conserving probability of this fine-tuning step in the CDR of the VH sequences is $FP_{CDR} = 0.46$ contrary to $FP_{FR} = 0.48$ in the FR. The average conserving probabilities of the VL_κ sequences are almost the same:

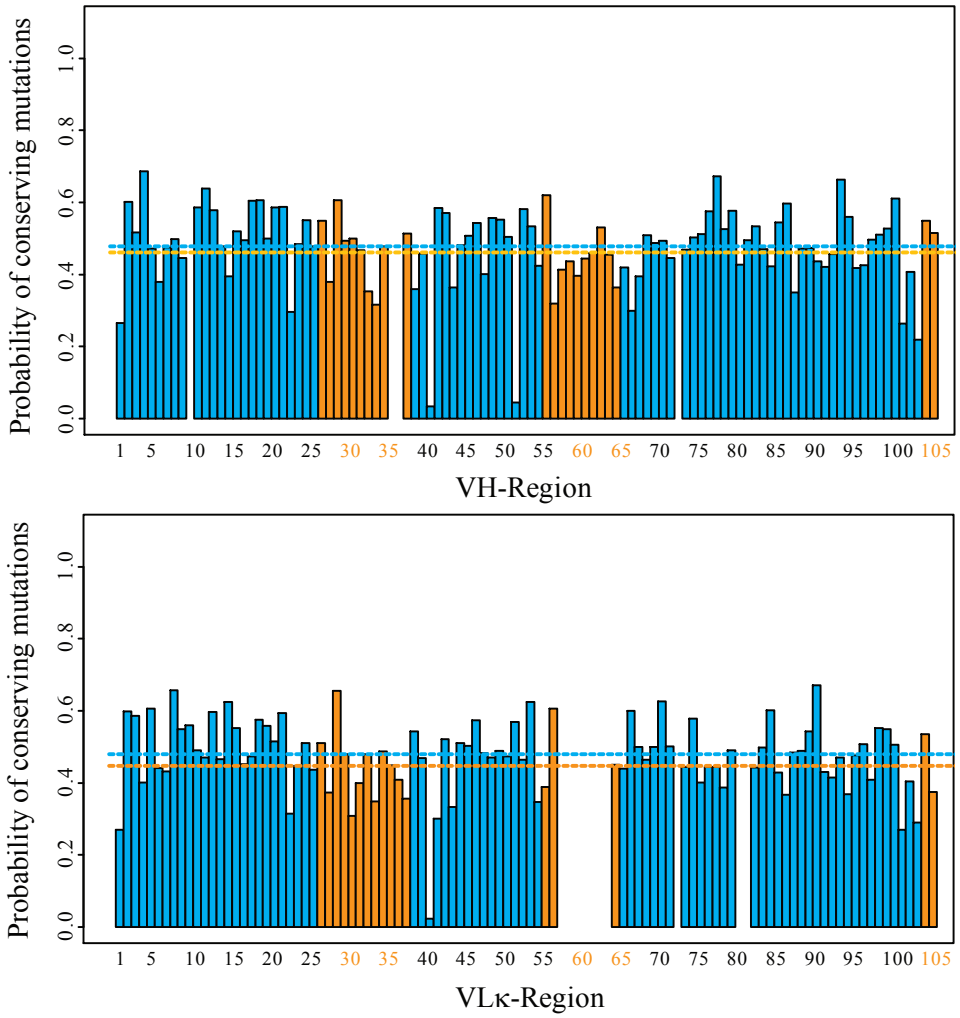


Figure 4.17: Potential single step mutation effects on the V genes of C57BL/6 mice. V gene sequences of C57BL/6 mice are aligned according to the IMGT format. The probability for a single point mutation (nucleotide based considering positional mutability as well) is computed for each codon according to AFFI[15] for all available V gene sequences. The higher the bars the more conserved is a mutation at that position on average. Orange colored bars denote mutations occurring within the CDRs, cyan bars depict mutations in the FR. The average conserving probability in these regions is depicted by the dashed lines. The gaps in the sequence are due to the chosen IMGT format.

$FP_{CDR} = 0.45$ and $FP_{FR} = 0.48$. The differences between CDR and FR for both V gene types is not significant according to a two-tailed Mann-Whitney U test ($p > 0.11$).

The conserved position W41 (codon TGG) of the VL_κ and also the VH sequences have a notably high probability for a harmful mutation. The codon TGG has no synonymous mutation in its neighborhood, but two STOP codons and the amino acids C, S, R and G, which have strong harmful probabilities to W. This indicates that each mutation at that position will result in a nonfunctioning receptor and thus emphasizes the importance of the conservation at that position. The VH sequences have a similar situation additionally at position 52, at which almost always the same amino acid appears: W encoded by its unique codon TGG.

Taken together, the coarse-tuning step of the immune system has high conservation in the FR vs. high diversity in the CDR, especially in the CDR2 of the heavy chains. It is notable that the average conservation probability in the CDR is more than twice as high as one would expect for a random mutation ($\text{AFFI}^{15}_{random} = 0.25$). Similar high conservation occurs for the fine-tuning step, which leads nearly to a 50% chance of conservation independent of the region. The high conserving probability here is mainly due to the "synonymous prone" neighborhood of the codons (Table 1.1). In fact, considering the transition/transversion bias (Equation (3.5)), 28% of all point mutations in the FR of both the heavy and light chain are on average synonymous. In the CDRs the synonymous mutations are decreasing, 26% in the heavy and 25% in the light chain, respectively. These differences are significant according to the significance range of a one-tailed Mann-Whitney U test ($p = 0.03$ for the heavy and $p = 0.004$ for the light chains). The decrease of the number of synonymous mutations benefiting the replacements in the CDR indicates a greater diversity there.

4.2.8 Discussion

A substitution matrix based on binding affinity only was generated and named AFFI. The entries of the matrix have a simple interpretation: they are the probabilities that an amino acid substitution conserves previous binding behavior. Note, that the data base used here consists solely of binary measurements of peptide binding affinities - more precisely of antibodies and their antigens in the form of linear epitopes. To be even more precise the deleterious substitutions are considered, because initially the antibody binds its wildtype antigen with high affinity. In contrast to data used in previous works for the computation of substitution matrices (BLOSUM, PAM), protein function, protein structure, foldability, the evolutionary career of pro-

teins, and overall amino acid frequency do not play a role in the establishment of AFFI.

Usually substitution matrices are given as log-odds matrices. The matrix entries are estimated by the logarithm of the fraction of the observed and the expected frequencies of a mutation. The advantage of the logarithm is the facilitated computation of sequence alignments (summing vs. multiplying scores). The great difference between those matrices and AFFI is the consideration of the expected frequencies. The big advantage of AFFI is that the underlying data basis comprises a complete substitution of each sequence position by any other amino acid. Therefore the whole mutational information for each position is available.

For further investigations the flexibility of the matrix was chosen to be 15. Obviously, a flexibility of 20 would include amino acid mutations that have at most backbone effects to the binding. With an undersized flexibility, deleterious mutations that cause binding damage would be excluded - only very closely related amino acids that have high probability of binding affinity conservation would be considered. A flexibility of 15 compromises both extremes. Furthermore, although $AFFI^5$ shows more predictive quality (Figure 4.9), it simultaneously has too many amino acid pairs that do not appear together in the dataset. This becomes clear when investigating the existence of neutral pathways of epitopes based on the key residues. The neutral pathways ensure connectivity between the amino acids. For $AFFI^5$ the connectivity of amino acids is low contrary to $AFFI^{15}$. Therefore $AFFI^5$ is not useful at the moment. There is a chance that for key residues with low flexibilities some amino acid pairs will never mutate into each other without binding loss - even if the size and diversity of the underlying dataset is larger. This uncertainty is the reason why another approach with a larger dataset has to be employed for determining which AFFI matrix is better for sequence alignment of e.g. CDRs. In any case, these newly defined matrices offer the opportunity to investigate residue similarities under new aspects.

The partition of amino acids was estimated with two different methods. Firstly, the proximity within the groups was maximized, and secondly the distance between the groups was maximized. As an additional parameter the groups were weighted according to their size. In principle all methods received similar results with slightly differences. Taking group sizes into account, influences almost only the partitions for similarity maximization. Group sizes have almost no effect on distance maximization. This is rather obvious as the weight is dependent on the number of interactions between groups and not directly on the group sizes. And the relative number of interactions doesn't vary much compared to the relative group sizes.

The optimized unweighted proximity clustering and both distance clus-

terings tend to produce small groups. The amino acids most distant from the others are excluded at first from the community. It is striking that the amino acids tryptophan and cysteine are the first ones to leave. Especially the cysteine could be expected here as it has the ability to form disulfide bonds, which are very important for stabilizing protein structures [128]. Tryptophan has several biologically specific characteristics: it fulfills apparently important functions in the structure of antibodies as indicated by its recurrence on defined positions in the FR. Besides methionine it is the only amino acid that is coded by only one triplet codon and additionally two out of the nine possible nucleotide point mutations within the triplet result in a STOP codon (Table 1.1). Another two mutations result in cysteine. This special composition of the genetic code around the tryptophan might be an evolutionary consequence of its obviously different binding behavior, and a reason for its importance in the FR of antibodies. The outstanding structural characteristic of tryptophan is that it contains an indole functional group. Besides the fact that aromatic amino acids can be involved in aromatic stacking interactions, this indole group makes the tryptophan special. It may be exposed to solvent [35]. Additionally, it makes tryptophan the most likely of the aromatics to be involved in a cation-π interaction [45]. However, besides the tendency to produce small groups with cysteine and tryptophan leaving first, the groupings of the optimized distance partitions are still similar to the weighted proximity partitions that are discussed in the following in more detail.

Residues with similar physicochemical properties are mostly grouped together, like the large hydrophobic and aliphatic residues (ILV) and the large and primarily hydrophobic aromatic residues (FYW), the longchained positively loaded basic residues (HKR), the small polar alcohols (ST) and the negatively loaded acidic residues (DE). The partition of the amino acids into two groups gives rise to the assumption that the separation of large (FHIKLMRVWY) and small (ACDEGNPQST) amino acids is preferred (Table A.7 - A.8). Previous investigations [21, 78, 96, 138] based on Blosum50, Blosum62 and Miyazawa-Jernigan (MJ) matrices [94] usually result in a classically HP model (a hydrophobic and a polar group). The large group might be classified as a hydrophobic group and the small one as the polar group, but the subgroup HKR usually has only few hydrophobic properties. There are also other investigations [80] not resulting in the HP-model, although based on the same matrices - Blosum50 and MJ. Therefore also the method chosen for clustering plays an important role here. The more groups are allowed in the optimized weighted proximity partitions the more the above mentioned groups become apparent. For $k = 10$ the first residue singling out is again tryptophan. For $k = 11$ the cysteine follows. Until the number of allowed

groups becomes too large, some residues with very similar chemical properties tend to stay together, such as the residues I and V, S and T, F and Y, K and R. These results are in accordance with others [78, 96].

Comparing the obtained representatives to other results gives less accordance. The five-letter alphabet is given as IAGEK [138], which agrees to the results found by others [80, 96] according to the experimental setup of Baker *et al.* [112]. Li found the letters YIGSE in a Monte Carlo approach based on the Blosum62 matrix [78]. While IAEGK does not include any aromatic amino acids, YIGSE does not contain any positively charged basic amino acid. But both amino acis groups seem to play an important role in antigen binding (Figure 4.8), especially the aromatic tyrosine or alternatively phenylalanine are found to be crucial [39]. These needs are considered in the reduced set presented here: SKIFQ (Table A.13).

Often there is the question asked, if there is a minimum number of residues required for a usable protein. Some experimental investigations answered that question positively [26, 106, 112]. What are the requirements on the characteristics of a protein?

1. it must fold into a stable and unique three-dimensional structure

2. the folding process must be realizable on an appropriate time scale

3. the protein must be able to perform its function

For the special case of antibody binding to its antigen the requirements are nearly concentrated on the function, as the folding is mainly solved by the FR of the antibody. The FR of both chains of all naturally occurring antibodies is significantly more conserved than the CDR (Figure 4.16) while the amino acid repertoire in the CDR is widespread spanning large parts of sequence space. That clearly shows the task sharing between FR and CDR. Additionally it could be shown that CDR codons are more prone to replacement mutations than FR codons, although the neighborhood of the codons is similarly conserved in both regions and in both chains. Interestingly, for human V genes Hershberg *et al.* [53] could show slightly different strategies to balance diversity and stability in the human immune response. In human VL_κ chains, the codons in both CDR and FW are more prone to replacement mutations, while in VH and VL_λ chains only CDRs are replacement prone - which is the same for mice VH and VL_κ genes. The small VL_λ chain contribution in mice to the repertoire of their variable light chains might be the reason for the fewer opportunities of their immune response, while in humans one light chain isotype (κ) enriches the immune response with more diversity even in the FR (and probably also more nonfunctioning).

Anyway, due to the clear task sharing between CDR (binding) and FR (structure) one might assume that the requirements on a reduced set of amino acids could be even smaller than for more complex proteins. For small proteins it was shown that five residues can be enough [112]. Murphy *et al.* [96] assume that 10 residues contain nearly as much information as all 20. The results presented here show that multiple single-point amino acid substitutions within $k = 6$ representatives have great conserving behavior. This again leads to the assumption that even fewer letters should have good opportunities to meet many binding criteria.

This awareness finally encourages to establish experiments based on reduced amino acid alphabets. For experimental sequence space search the combinatorial diversity of peptides is too large. Even a peptide with a length of seven amino acids - which covers most of the investigated epitopes [30] - has more than 10^9 possible amino acid combinations. By using a reduced set of only five or six amino acids the size of the sequence space would decrease tremendously to less than 10^6 peptides. The today's technical opportunities like the SPOT synthesis on cellulose sheets or even more modern techniques like synthesis of peptides on glass slides allow to synthesize that masses.

In the results presented here a reduced set of amino acids consists of one amino acid out of each group. The reduced set derived from p_5 in Equation (4.19) is shown to be the best way to span the sequence space - and it is more promising than using random libraries. An alternative of choosing the representatives is to choose the optimal partition according p_1 in Equation (4.15), which optimizes the proximity of the group members while ignoring the group sizes. This partition has the disadvantage that one group results in a remainder group, collecting all amino acids that do not nicely fit to the others. This means that the group members of the remainder group do not necessarily have many similarities. The advantage on the other hand is that all other groups are very strongly correlated. Now, the representatives that could be chosen are derived from the strongly correlated groups while ignoring the remainder group. The benefit here is that on the one hand the part of the sequence space covered by the strong groups is covered very well, while on the other hand the probably even bigger part of the sequence space is not covered at all or even only barely.

In conclusion, the decision which cluster is the best basis for the choice of the representative amino acids depends on the problem and if there is some previous knowledge on the respective epitope, regarding its length or its physicochemical properties. If there is nothing known, it is suggested to choose an optimized partition obtained via p_5 in Equation (4.19), which spans the sequence space most uniformly and chooses the best representative out of each group simultaneously.

An adjacent application area of the substitution matrix AFFI is the investigation of the output of germinal centers as already introduced above. This output will be analyzed in more detail in the next section.

4.3 Recurrent mutations in one canonical antibody heavy chain sequence from mice

4.3.1 Extraction and selection of $VH_{186.2}$ sequences

One essential point of this part of the thesis was to collect the largest possible, representative yet reliable pool of $VH_{186.2}$ sequences. B cells bearing these gene sequences for BCR production are almost always chosen for a germinal center reaction by the immune system of the C57BL/6 mice when infected with NP. This fact allows to investigate mutations that do really have impact on the affinity maturation of antibodies. One very important precondition here is the clonally independence of the selected sequences. Any distortion of the assessed frequency distribution of mutations due to clonal redundancy of mutations was avoided by applying strict selection criteria for $VH_{186.2}$ sequences, as specified in the Materials and Methods. Moreover, the collected data includes sequences derived from both different lymphoid tissues and different time points after primary NP-CGG challenge (for details, see Table 3.2). The resultant pool of $VH_{186.2}$ sequences constitutes the data basis for the analysis presented here and is specified in Table 3.2. For reasons of simplicity, $VH_{186.2}$ sequences were assigned to three time intervals, namely early phase (day 6 to 10), peak phase (day 11-21) and late phase (day 30 to 140) of the primary immune response. In total, the dataset includes 781 $VH_{186.2}$ sequences, of which 224 sequences belong to the early, 321 to the peak and 208 to the late phase. A further 28 sequences could not be assigned to any of the three phases because the time point of isolation was not given. The majority of sequences ($\sim$ 90%) were isolated from lymphoid tissues such as spleen, lymph nodes and nasal-associated lymphoid tissue. The remainder were isolated from bone marrow. Out of the 781 $VH_{186.2}$ sequences, 281 ($\sim$ 36%) were identified as germline sequences. Germline $VH_{186.2}$ sequences were found in the early, peak and late phase, with decreasing frequency (38%, 36% and 27%, respectively) (Figure 4.18). The number of mutations identified per $VH_{186.2}$ sequence ranged between 0 and 23. The average mutation frequency was shown to increase over the course of time (2.67, 2.97 and 5.09, respectively), and the distribution of the numbers of mutations shifted towards higher numbers (Figure 4.18). In addition, the fraction of the well-described W34L (W33L, according to Kabat nomenclature [61]) key mutation turned out to increase consistently throughout the three predefined phases (5%, 12% and 17%). The observed time-dependent accumulation of mutations, in particular the accumulation of the W34L key mutation, reflects ongoing affinity maturation.

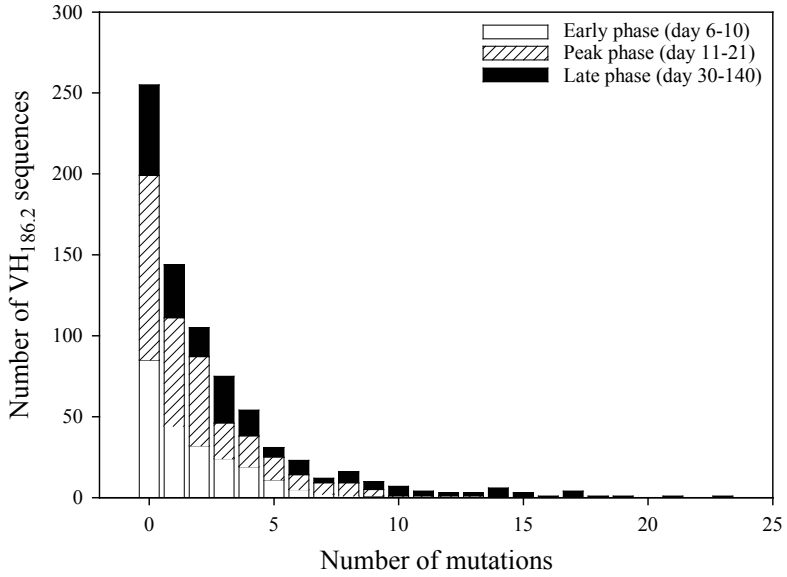


Figure 4.18: Distribution of the number of mutations in collected $VH_{186.2}$ sequences recovered during the primary NP response. The stacked bar graph illustrates the distribution of the number of mutated $VH_{186.2}$ sequences obtained from C57BL/6 mice during the early, peak and late phase of an immune response against NP (for details, see Table 3.2).

4.3.2 Predicted relative frequency distribution of amino acid substitutions

The analysis revealed a total number of 684 potential amino acid substitutions in the FR1 to FR3 region of the $VH_{186.2}$ chain. This count includes not only nonsynonymous but also synonymous amino acid substitutions. For the sake of simplicity, the term "amino acid substitution" will be used to denote both synonymous and nonsynonymous substitutions.

The number of possible amino acid substitutions per codon ranges between six and eight, with the average being seven. To address the predicted incidence of these substitutions, first the expected frequency of amino acid substitutions defined by the intrinsic mutability of the $VH_{186.2}$ chain is assessed. Two aspects of the intrinsic mutability of the $VH_{186.2}$ chain are considered:

- the probability of a given nucleotide to mutate depending on its sequence context, which is referred to as codon positional mutability throughout the thesis,
- the chance that a nucleotide mutates into the other three nucleotides as specified by the well-known transition bias of mutations [10, 137].

Thus, the predicted frequency of amino acid substitutions comprises the codon positional mutability f_c (height of bars in Figure 4.19) and the likelihood that mutations within a given codon result in a certain amino acid (height of subbars in Figure 4.19). The predicted frequency distribution shows no systematic pattern. However, two segments of rather low mutability stand out. Both are localized at transition regions, namely FR2/CDR2 (positions $45 - 58$) and FR3/CDR3 (positions $94 - 101$).

4.3.3 Observed frequency distribution of amino acid substitutions

The observed frequency distribution of amino acid substitutions revealed by analysis of all 781 $VH_{186.2}$ sequences (Table 3.2) is illustrated in Figure 4.20. Again, codons of both high and low mutability were identified within regions encoding FRs and CDRs. The codons 74, 66, 34 and 32 are found to have the highest mutability.

The observed mutations may also be computed and therefore judged with respect to conserving probability with the substitution matrix AFFI[15] (Table A.3). For each codon of the $VH_{186.2}$ chain the conserving probability of the occurring mutations are averaged resulting in Figure 4.21. The conserving probability of the mutations in the CDRs is on average 0.49, while it is 0.57 in the FRs. The expected average conserving probabilities of the fine-tuning step (Equation (4.24)) for the $VH_{186.2}$ sequence alone are $FP_{CDR} = 0.44$ and $FP_{FR} = 0.47$, respectively. The real mutations have greater conserving probabilities than the theoretically predicted ones, which is significant for the FR (one-tailed Mann-Whitney U test: $p = 0.001$) and not significant for the CDR (two-tailed Mann-Whitney U test: $p = 0.50$). There is a tendency for conservation within the FR during the investigated germinal center reaction.

4.3.4 Incidence of favored amino acid substitutions

The amino acid substitutions in the dataset of $VH_{186.2}$ sequences rated as favored, are illustrated in Figure 4.22. Out of the 684 possible substitutions in the FR1 to FR3 region of the $VH_{186.2}$ chain (Figure 4.19), 23 substitutions were rated as favored, that is about 3% of all potential substitutions. Indeed, a total of 23 favored substitutions were identified at 20 different positions of the 96 amino acid long FR1 to FR3 region (codons 1 to 104) of the $VH_{186.2}$ chain. Accordingly, about every third position showed either one or more favored substitutions. However, the incidence of positions featuring favored substitutions is four times as high in CDR (0.56) as in FR (0.14).

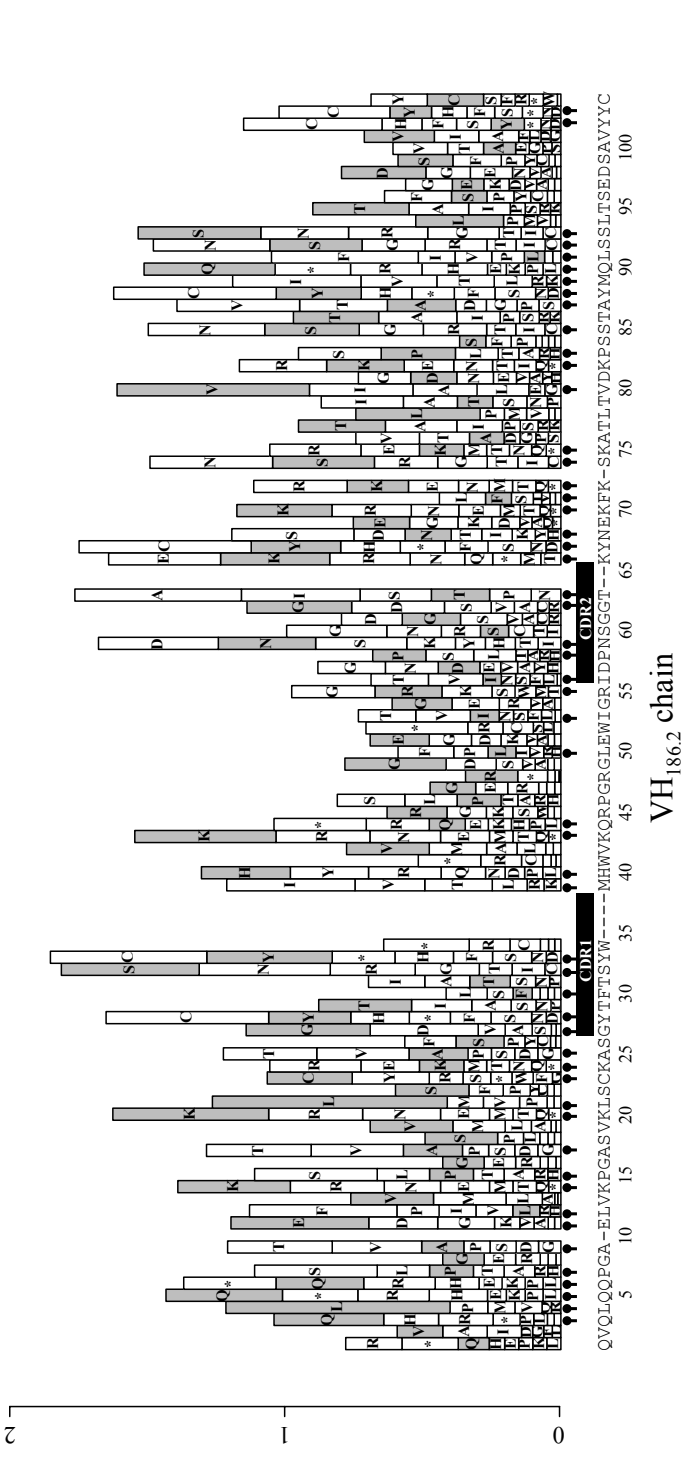


Figure 4.19: Predicted amino acid substitution frequency distribution of the $VH_{186.2}$ chain. The predicted relative codon positional mutabilities f_c reflect the tendencies of codons to mutate. A value of 1 represents the average mutability of codons in the $VH_{186.2}$ chain. Each subbar denotes the relative substitution frequencies into the indicated amino acids with respect to the well-defined transition bias of mutations [10]. The germline amino acid sequence of the $VH_{186.2}$ chain is displayed according to the IMGT unique numbering system [77]. Regions corresponding to CDRs are indicated. For comparison purposes, synonymous amino acid substitutions are highlighted gray, and hotspots of the $VH_{186.2}$ chain estimated according to [114] are marked above the $VH_{186.2}$ amino acid sequence.

Figure 4.20: Observed amino acid substitution frequency distribution of the $VH_{186.2}$ chain. The observed codon positional mutabilities as revealed by analysis of 781 $VH_{186.2}$ chain sequences are shown (for details, see Table 3.2). Each subbar denotes the according relative amino acid substitution frequencies. Only frequencies with more than five mutations are displayed. The germline amino acid sequence of the $VH_{186.2}$ chain is shown and annotated as described in Figure 4.19.

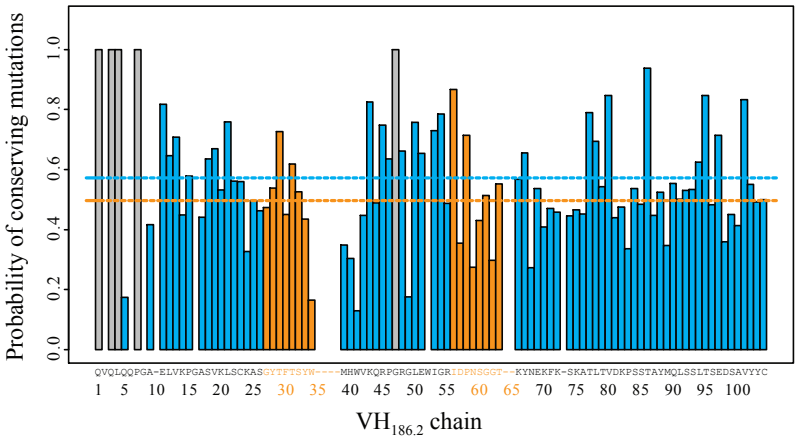


Figure 4.21: Observed amino acid substitution of the $VH_{186.2}$ chain viewed with AFFI. The conserving probabilities of the observed mutations within the FR1 to FR3 region of the $VH_{186.2}$ chain sequences according to the substitution matrix AFFI[15] are shown. The gray bars denote that on the respective position only synonymous mutations take place, while the orange bars denote mutations within the CDRs, cyan bars are mutations in the FR. The average conserving probability in these regions is depicted by the dashed lines.

From the 23 substitutions rated as favored, four are synonymous, another eight are substitutions by amino acids showing similar physical properties and AFFI[15] (Table A.3) shows high conserving probabilities, e.g. substitution of polar and small (S32T, S60T, S74T, S85T) and acidic residues (E11D). On the other hand, the key mutation W34L comprises a dissimilar amino acid change according to AFFI[15] ($p = 0.17$). Only three further replacements have a conserving probability of less than its average ($p = 0.25$): S32N ($p = 0.15$), S74N ($p = 0.15$) and N59Y ($p = 0.10$). The position of the serine to asparagine mutation has in both cases additionally a favored serine to threonine mutation. Comparison of the properties of these three amino acids (Figure 1.5) demonstrates that they are all small, but serine is classified as tiny.

Regarding the reliability of the chosen model, the incidence of favored amino acid substitutions was shown not to depend on realistic variations of the parameters for the intrinsic mutability of the $VH_{186.2}$ gene. Both, assuming a homogeneous positional mutability by ignoring the effects of hotspots ($f_i = 1$) or by considering each position to be a hotspot ($f_{i,max} = 2.81$), and including variations of the transition/transversion bias in a range from $1:4$ to $4:1$ lead only to minor changes regarding the pattern of detected favored substitutions. The ten most frequent favored substitutions and the statistical significance of the findings are not changed by any of these manipulations.

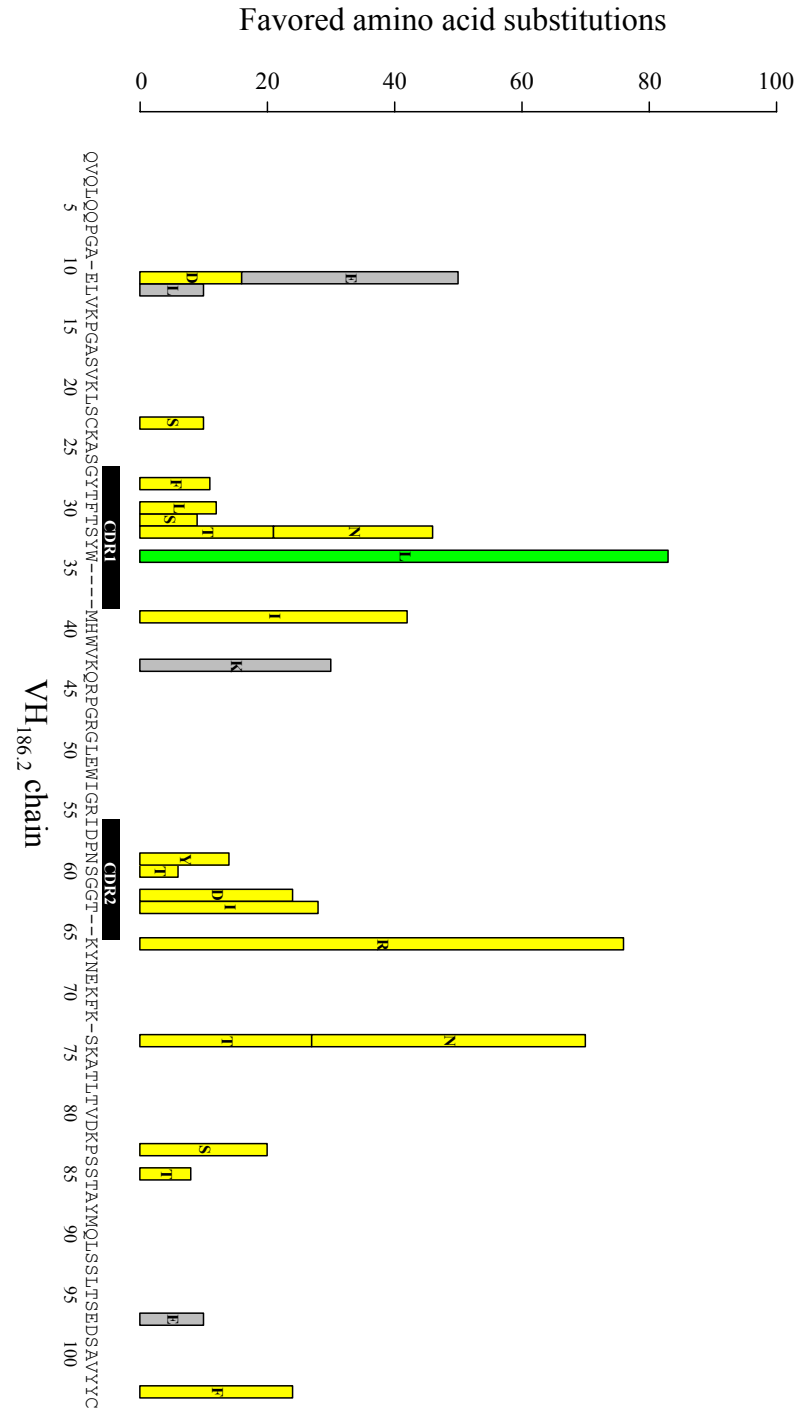


Figure 4.22: Incidence of favored amino acid substitutions in the $VH_{186.2}$ chain. Favored substitutions were identified by statistical comparison of the predicted and observed frequency distributions of point mutations as described in Materials and Methods. Substitutions rated as favored are colored yellow, the key substitution W34L is distinguished in green. Favored synonymous substitutions are highlighted in gray. The y-axis gives the absolute number of $VH_{186.2}$ chain sequences bearing the respective substitution. The germline amino acid sequence of the $VH_{186.2}$ chain is shown and annotated as described in Figure 4.19.

Ranking	Favored AA substitution	Frequency [%][a]	Region
1	W34L	16.9(10.8)	CDR1
2	K66R	15.2(9.7)	FR3
3	S74N	8.7(5.6)	FR3
4	M39I	8.3(5.3)	FR2
5	**E11E**	6.9(4.4)	FR1
6	**K43K**	5.9(3.8)	FR2
7	T63I	5.6(3.6)	CDR2
8	S74T	5.5(3.5)	FR3
9	S32N	5.2(3.3)	CDR1
10	Y103F	4.8(3.1)	FR3

[a]Frequency of $VH_{186.2}$ chains carrying the indicated favored amino acid substitutions within the dataset excluding germline sequences; values in brackets indicate the frequency when germline sequences are included ($N = 781$). Synonymous mutations are marked bold.

Table 4.4: Ranking of favored amino acid substitutions

4.3.5 Frequency of favored amino acid substitutions

Among the sequences (FR1 to FR3) showing amino acid substitutions ($n = 139, 207, 152$, in the early, peak and late phase), the proportion of those carrying favored ones increases steadily throughout the three phases (51%, 61%, 76%). However, the distribution of the relative frequencies of favored substitutions in the respective $VH_{186.2}$ chains was rather invariant with respect to time (Figure 4.23) and distributions proved to be identical according to the significance range of two-tailed Mann-Whitney U tests ($p > 0.24$). Regardless of the phase of the immune response, nearly every second substitution in the dataset of sequences containing favored substitutions turned out to be a favored one (41%, 45%, 39%). Considering all sequences the relative frequency of favored substitutions is about one third (26%, 35%, 35%).

A ranking of the ten most frequent substitutions is given in Table 4.4. The well-described W34L key substitution in CDR1, found to be most frequent, was identified in 84 of the collected $VH_{186.2}$ chain sequences (11%). A K66R (K58R according to Kabat nomenclature [61]) substitution in FR3 was found with a similar high incidence; it was identified in 76 $VH_{186.2}$ chain sequences (10%). The K66R replacement was previously described as being recurrent and proposed to constitute a second key mutation of the anti-NP response [43, 85]. As mentioned above for the W34L key substitution (Table 3.2), the proportion of K66R bearing $VH_{186.2}$ chains also increases from the early to the late phase (5%, 11% and 14%, respectively). Furthermore, substitutions at position 74 of FR3 (S74N/T) have similar summed frequencies ($6 + 3 = 9\%$). Notably, out of the six most frequent substitutions, two turned out to be

synonymous (E11E and K43K).

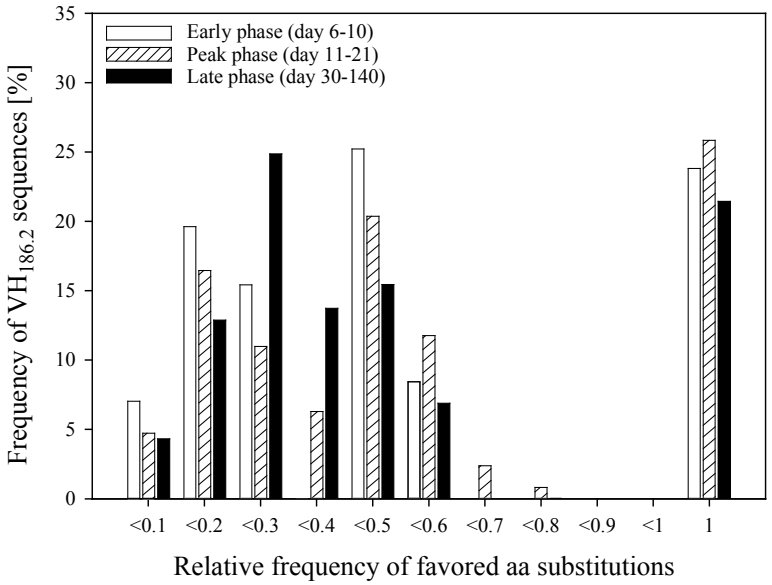


Figure 4.23: Fraction of favored amino acid substitutions in $VH_{186.2}$ chains. These fractions were recovered in the early ($n = 71$), peak ($n = 127$) and late phase ($n = 116$) of the primary NP response. Relative frequencies are recorded as the ratio of the number of favored and the total number of substitutions found in every $VH_{186.2}$ chain of the dataset carrying favored substitutions.

4.3.6 Localization of favored amino acid substitutions in the 3-D structure

To localize relevant amino acid substitutions, 3-D visualization of a $VH_{186.2}$/ $VL_{\lambda 1}$ Fv fragment complexed with a NP compound was performed. For the sake of clarity, the image of the 3-D structure is confined to the $VH_{186.2}$ chain and the NP compound (Figure 4.24). The two top-ranking substitutions W34L and K66R, as well as the fourth-ranking M39I substitution (Table 4.4), are shown to be close to the binding pocket region of the $VH_{186.2}$ chain. The seventh-ranking T63I substitution is adjacent to the K66R substitution. The remaining relevant amino acid substitutions (S32N, S74N/T, Y103F), including the favored synonymous substitutions K43K and E11E, are located in very distinct sites distal to the binding pocket region of the $VH_{186.2}$ chain.

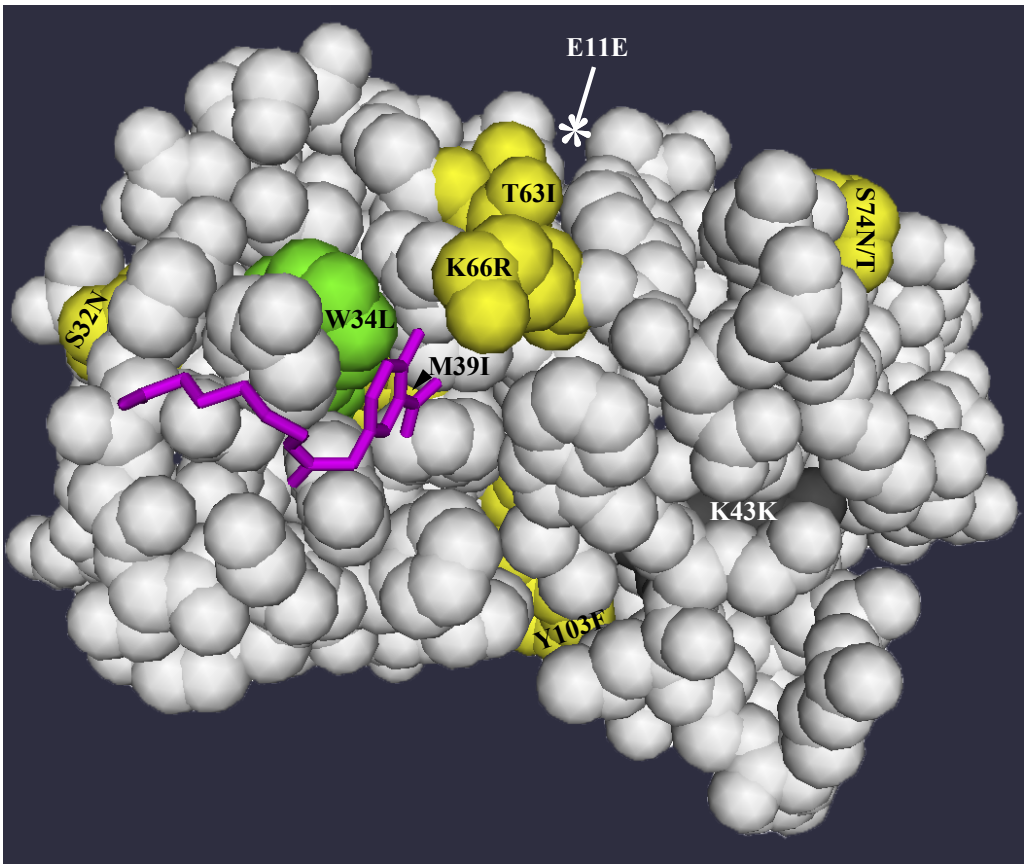


Figure 4.24: Relevant mutations are located at sites both nearby and distal to the binding pocket. The 3-D structure of a $VH_{186.2}/VL_{\lambda 1}$ Fv fragment complexed with a NP compound obtained from the Protein Data Bank [70] (PDB ID: 1a6v) was visualized as a Corey-Pauling-Koltun space-filling model using the molecular visualization system PyMol [28]. The $VL_{\lambda 1}$ chain was masked out in the illustration for clarity. Indicated are the top ten ranking favored amino acid substitutions (Table 4.4). Substitutions rated as favored are colored yellow, the key substitution W34L is distinguished in green. Favored synonymous substitutions are highlighted in gray. One synonymous substitution only visible in the back view of the 3-D structure is indicated by an asterisk. The NP compound is indicated in purple.

4.3.7 Assessment of signatures of antigenic selection

Antigenic selection is usually assessed by comparing the observed frequency of replacement (R) and silent (S) mutations to the expected frequencies under the null hypothesis of no selection [9, 69]. Here, the focused binomial test recently published by Hershberg *et al.* [54] was applied. The observed frequencies of favored R mutations in both CDRs and FRs were shown to exceed the expected value (E_0); the differences turned out to be statistically significant (Table 4.5). That is, favored mutations in CDRs and FRs show signatures of strong positive antigenic selection. This is still the case when the W34L CDR substitution is excluded from the statistical test, while exclusion of the K66R mutation avoids significance in the FR. By contrast, the observed frequencies of non-favored R mutations in CDRs and FRs were below the expected value (Table 4.5). The statistical test further reveals that non-favored mutations in CDRs and FRs do not show signs of antigenic selection. Taken together, the collected data indicates strong positive selection of favored mutations, in particular of those located within CDRs.

	CDR	**FR**
Expected frequency (E_0)[a]	0.37	0.68
Observed frequency[b]		
Favored	$0.74^{++}(0.64^{++})$	$0.76^{++}(0.69)$
Non-favored	0.37	0.66

[a]Expected frequencies (E_0) under the null hypothesis of no selection (P_0) calculated as the fraction of R mutations within the region of interest (FR or CDR) and the overall number of mutations ($R+S$) within the $VH_{186.2}$ chain according to the focused binomial test developed by Hershberg *et al.* [54].
[b]Observed frequencies are given as the fraction of R mutations within the region of interest (FR or CDR) and the overall number of mutations ($R+S$). $VH_{186.2}$ chain sequences were rated to show positive (+) or negative (−) antigenic selection for $p < 0.05$ and strong antigenic selection (++, −−) for $p < 0.01$, respectively. Values in brackets indicate observed frequencies when the key residues W34L and K66R were excluded.

Table 4.5: Assessment of signatures of antigenic selection of favored and non-favored mutations in the $VH_{186.2}$ chain during the primary immune response to NP.

4.3.8 Discussion

The mutation spectrum of antibodies due to somatic hypermutation is a combined result of both the intrinsic ("background") mutability of their VH/VL chain sequences and the effects of selection. This situation particularly complicates assessment of the selection process itself and identification of associated signatures of selection. Here, new aspects of the diversification and selection of $VH_{186.2}$ sequences during the primary immune response to NP

are reported revealed by statistical comparison of expected and empirically observed frequency distributions of mutations. Based on a large empirical library of $VH_{186.2}$ sequences, such analysis allowed the identification of mutations that occur more often ("favored") than expected.

The resulting favored mutations and related amino acid substitutions turned out to be very robust against variations of the chosen model parameters. Notably, the ranking of the ten most frequent substitutions (Table 4.4) remained unchanged with regard to all applied variations. The variations in model parameters included changes in the intrinsic mutability of the $VH_{186.2}$ gene, namely in the positional mutability f_i and in the probabilities for transitional and transversional point mutations (p_m). Additionally, an alternative estimation of the empirical mutation probability (p_0) computed without the inclusion of non-mutated sequences did not essentially alter the obtained results.

The presented model does not take into account long range effects caused by mutations at hotspots, as described by Clark *et al.* [22]. However, these effects never lead to neighboring positional mutabilities that exceed the mutability of the hotspot. Therefore, the variation of the model parameters by which each position is regarded to be a hotspot ($f_{i,max} = 2.81$) already covers long range effects, showing that disregard of these effects does not lead to essential changes in the ranking of favored substitutions.

In conclusion, the accumulation of favored substitutions can only be due to background mutability if the corresponding positional mutabilities reach peak values not yet reported in literature. Comparison with a negative control dataset could provide further evidence as to whether mutability values at yet unreported heights at around 20% of the positions are the explanation for the large number of substitutions rated as favored. This being unlikely the recurrence of these mutations could be a consequence of ongoing selection.

It is difficult to assess if selectively neutral mutations accumulate in the pool of favored mutations. The potential accumulation of selectively neutral mutations is directly linked to the as yet unknown ratio of mutations. If the vast majority of mutations is deleterious, as suggested by Shlomchik *et al.* [123, 125], selectively neutral mutations are likely to accumulate in the pool of favored mutations. In contrast, a high number of selectively beneficial mutations would rather support accumulation of neutral mutations within the pool of suppressed. So far, both the identification of only particular synonymous mutations within the pool of favored mutations and the signatures of positive selection of favored mutations suggest that the mutations rated as favored were not selectively neutral. In the case of S32N/T and S74N/T the mutant amino acids share the same property of being slightly larger than the original one indicating a directed mutation providing evidence for positive

selection. The relevance of the identified favored mutations is further emphasized by the fact that in particular those found in CDRs show signatures of strong positive antigenic selection (Table 4.5).

Most notably, the number of mutations identified as favored was unexpectedly high ($n = 23$), indicating that affinity maturation of the $VH_{186.2}$ chain involves not only a few but a whole spectrum of relevant mutations. Here, the familiar W34L key mutation and also the K66R replacement proved to be the most frequent favored mutations. Previously described as being recurrent, the K66R replacement is thought to represent a second key mutation of the anti-NP response [43, 85]. Whereas the favored key mutation positions at codons 34 and 66 were each replaced by one specific amino acid, other codons such as 32 and 74 were shown to include two different favored substitutions (S32N/T and S74N/T). In this context, the overall positional mutability of codon 74 turned out to be about equal to those of W34L and K66R. The fact that "mutating away" from an amino acid might correspond to various different substitutions helps to explain why this high mutability positions were overlooked in previous investigations.

Discovering a whole spectrum of recurrent and therefore potentially relevant mutations demands careful reconsideration of the term "key mutation". Since affinity measurement data for the newly identified favored mutations is lacking, it is still open as to whether the term "key mutation" only applies to W34L, or if further mutations strongly contribute to increasing affinity of the $VH_{186.2}$ chain. The presence of a broad spectrum of relevant mutations and their continuous accumulation suggests that affinity maturation might be due to the sum of many small effects. Improvement of binding characteristics through cumulative effects of many small structural alterations has already been reported for antibodies directed against the hapten fluorescein [90].

In the investigated dataset of sequences containing favored substitutions, nearly every second mutation among those observed is a favored one. This ratio remains constant in the early, peak and late phases of the response (Figure 4.23), showing that, throughout the response, favored and selectively neutral mutations accumulate almost with the same rate. This finding indicates that affinity maturation is still ongoing in the late phase of the immune response.

Various affinity maturation studies have shown that mutations generally occur in both antigen contact and noncontact regions [20, 25, 79, 134, 140]. Interpretation of this finding is hindered by the fact that the reported mutations' contribution to an increase in affinity is not known. However, the results presented here reveal for the first time that mutations found in both antigen contact and noncontact regions are favored by selection. Although preferential localization of favored mutations in CDRs forming the binding

site was observed, the 3-D representation of a $VH_{186.2}/VL_{\lambda 1}$ Fv fragment subsequently demonstrated that most of the favored mutations are located at sites distal to the binding pocket (Figure 4.24). This lends support to the concept that improvement (fine-tuning) of antibody binding characteristics involves optimization of the binding pocket periphery, as proposed by Li *et al.* [79]. Indeed, such optimization of the periphery need not necessarily affect affinity, but may involve other characteristics, such as folding efficiency and resistance to proteolysis. This implies that selection acts on other factors besides affinity, as previously proposed by Shlomchik *et al.* [124].

A rather striking finding is the apparent importance of several synonymous mutations for affinity maturation of the $VH_{186.2}$ chain. Thus, selection might not only act at the level of the protein, but already at the level of the nucleotide sequence (mRNA). The pool of favored mutations includes four synonymous mutations, of which two (E11E, K43K) ranked fifth and sixth in the frequency of occurrence (Table 4.4). The fact that particular synonymous amino acid substitutions are indeed recurrent and attributable to independent processes is further emphasized by the fact that these mutations were found by different laboratories. The two most frequent synonymous amino acid substitutions, E11E and K43K, were each found in around 50% of all mice. At first glance, it seems surprising that affinity maturation involves accumulation of synonymous mutations, since affinity is certainly not directly enhanced by this type of mutation. But, positive selection of synonymous mutations has been shown to be common among mammalian genes. In fact, it is even more common than positive selection at the level of protein sequence [110]. Positive selection in synonymous sites has been linked to the stability of mRNA secondary structure, thereby affecting gene expression and translation rates [16, 31, 117].

Moreover, selection at synonymous sites might also be related to tRNA adaptation by favoring those codons that match the most abundant tRNAs [1, 62]. In addition, positive selection of synonymous mutations could be driven by the need to facilitate interactions between the (Ig) mRNA and small regulatory RNAs [88]. Indeed, it could be shown that introduction of an unpreferred synonymous codon into the Drososphila *Adh* gene results in reduced levels of the ADH protein [14]. These effects have not been studied in B cells so far. However, GC B cells have a rather short life time, that is, they have to rapidly produce adequate amounts of BCRs before undergoing selection.

This strongly suggests that achievement of fast and efficient gene expression rates may even play a more important role for BCRs than other proteins, and it is straight forward to assume that the very same mechanism that leads to affinity maturation also enforces “expression rate maturation”. This idea

is supported by the fact that BCR sequences bear random nucleotide insertions and deletions, and hence certainly are not evolutionarily tuned towards optimal expression rates.

The analysis further reveals that all but one of the collected $VH_{186.2}$ chain sequences carry either a W34L or a K66R substitution or none of both. This result is in accordance with previous reports by Furukawa *et al.*, who proposed different binding modes of NP and the existence of two different maturation pathways for anti-NP antibodies [43, 44, 85]. However, sequences carrying either W34L or the K66R substitution share a great number of favored substitutions (S32T/N, M39I, S74T/N) including the favored synonymous substitutions E11E and K43K. This finding lends support to a scenario in which division into the two groups occurs rather late in the response - after accumulation of the shared favored mutations. Admittedly, this scenario is hypothetical, since no information is available regarding the chronological order of acquiring mutations. Nevertheless, it provides a possible explanation for the particular long delays before the appearance of key mutations as first pointed out by Radmacher *et al.* [68, 103]. In this context, one might speculate that affinity maturation first involves a preliminary phase, which increases the non-affinity related fitness of B cells. In view of the accumulation of shared favored mutations, in particular the synonymous ones, the aim of this phase could be the attainment of sufficient and stable surface BCR expression levels, e.g. through altering mRNA stability and secondary structure, resistance to proteolysis and protein folding (as argued before).

The limited number of sequences available sets a limit for the identification of the observed mutations. Some of the mutations, not rated as favored in this work, may well be favored ones, albeit to a lesser degree. Furthermore, the presented approach can be extended to identify negatively selected mutations - mutations occurring less often than expected. However, the insufficient size of the currently available dataset ($n = 781$) allows only for a partial identification of suppressed mutations. Identification of all potentially suppressed mutations would require a dataset of far more than $20,000$ $VH_{186.2}$ sequences, which is currently beyond reach.

Chapter 5

Conclusions

The main objective of this thesis was to shed light on the functional relationship of amino acids and how this relationship is connected to peptide-antibody binding affinity changes, and moreover, to the antibody affinity maturation process. The solid work of Liying Dong [30] offered the opportunity to investigate the amino acid network that is responsible for the binding affinity of antibodies. To make her data powerful for quantitative evaluation of the relationship between amino acids, the reliability of the experimental background was investigated first.

Experimental background

A method to reduce the standard deviation of signals measured with the SPOT technique was described. It is able to diminish the standard deviation from as much as 22% of the mean value down to 13% (Figure 4.3A). To make the best use of this method, it is suggested that around 4% of the spots on a membrane should be homogeneously distributed repeats of a reference peptide possessing high affinity towards the ligand, if available.

This leads to

1. optimal results in the algorithm for reducing signal noise due to regional trends; and

2. optimal conditions for automatic grid positioning by the software that measures signal intensities.

For experiments where no strong binding peptides are known beforehand, the noise reduction algorithm can still be applied with only background signals as references. In this case noise reduction is suboptimal (Figure 4.3A).

There is good agreement between model results based on the mass action law (Figure 4.4A). The model takes into account competition among

peptides for free antibody molecules. It describes several observed effects, such as a shift in the dynamic range and hence the border separating high and low affinity classes, when only changing the number of peptide replicas on a membrane (Figure 4.4B). Furthermore, the model takes into account the apparently contradictory behavior of different peptides' SIs when comparing experiments performed with different amino-functionalization (FQ). This observation suggests that competition effects do indeed play a role in the usual experimental setup, and that, due to this effect, inter-membrane SI comparisons have to be handled with care even after normalization of the values (but note, this effect does not disrupt the ability to classify peptides according to their affinity when the SIs used were measured on a single membrane). It also suggests that a decrease in effective peptide concentration, possibly by dramatically decreasing FQ, could help diminish this effect.

Fitting of the model to experimental data in order to estimate experimental parameters suggests that the effective concentration of accessible peptides on the cellulose membrane decreases for increasing FQ. It is expected that for an FQ much smaller than 10%, this behavior can be reversed. Moreover, the fitted parameters suggest that the effective ligand concentration within the membrane fibers increases sub-linearly with increased ligand concentration in the bulk solution.

Classification of peptides into two classes representing high and low binding affinities is possible with high predictive power using the measured signal intensities as criteria. The SI-threshold, separating both classes, is best chosen as the mean value of the SIs of the spot background signals plus three times its standard deviation. Recently, the method derived here was successfully applied to characterizing PDZ domain/ligand specificity [11, 143].

The SPOT technique also allows differentiation between three classes of dissociation constants: high, low and intermediate. The most frequently used experimental condition ($FQ = 50\%$ and $[A_T] = 6.3\,nM$)) yields an intermediate class whose affinities lie between $pK^1_{dis} = 5.3$ and $pK^2_{dis} = 6.8$. This dynamic range, and with that, the pK^1_{dis} border of classification in two classes can be shifted by altering the effective concentrations of the binding partners. There are two possibilities for varying the peptide concentration: changing the FQ or using a differently coated membrane, e.g., CAPE, as described in [136]. Furthermore, the ligand concentration in the bulk solution can be varied.

In another work, the results presented here could be expanded to the more modern high throughput technique of printing peptides on glass slides [133]. The results obtained were and surely will be incorporated into many other investigations and activities of different workgroups. Furthermore, they allow for the first time, an analysis of even older measurement data in an

objective way, so that new results may be generated. A case in point was the re-analysis of the 10 year old data from Liying Dong's thesis [30].

Substitution matrix AFFI

The data analysis tools provided here, permitted the creation of the first substitution matrix that was based on binding affinity data only, named AFFI. The comparison of the binding affinity prediction quality of this matrix with the very common BLOSUM and PAM matrices gave better results for AFFI.

The partition of amino acids into distinct groups that share common properties in the sense of binding affinity based on AFFI has striking analogy to partitions based on other matrices that have protein families as background, which includes amino acid similarities that are important for folding. It is noteworthy that the properties of amino acids that are responsible for folding and that for binding seem to be the same.

Epitope identification has used small antigen sequence-derived peptide libraries (peptide scans), huge libraries of individual peptides prepared biologically, combinatorial libraries with peptide mixtures and random libraries with a small set of peptides. The amino acid partition was used to derive a reduced set of amino acids, that gives hope to be a perfect fundament for epitope search. This hope was supported by a theoretical approach presented here. Further support is seen in the finding that most selected replacement mutations were of conserving nature (Figure 4.22), so that even for larger Hamming distances (which are unavoidable here) the conserving probability is high (Figure 4.13). Additionally, it must be stated that an optimal reduced set of amino acids might play a substantial role in many fields of current interest. This is especially true for the search of new medicaments and also for diagnostic tools.

Some preliminary experimental tests yielded promising results. A starting peptide array of less than 6000 individual 20-mer peptides prepared by SPOT synthesis, based on a reduced set of amino acids and optimal library design as proposed in this work was used to successfully identify peptides that bind to the model antibody CB4-1. Comparing the number of identified epitopes with other approaches is not straightforward; the different approaches use different detection methods and simple comparison of the quantity of epitopes does not offer much help. In any case, some non-homolog epitopes have been found, which had dissociation constants in the range of $pK_{dis} = 5$ to $pK_{dis} = 6$. Further substitution analysis of binding peptides from the starting peptide array revealed the critical residues important for binding. Using this information and all 20 amino acids for one further optimization

step results in even higher binding affinities ($pK_{dis} = 7$). Some even smaller optimized peptide libraries based on a smaller reduced set incubated with other antibodies as well as proteins also succeeded in binding.

There are still some open questions regarding this topic: On which kinds of proteins can a reduced set of amino acids successfully act? Are there different optimal reduced sets for different purposes? The development of a tool would be desirable for the quality control of the results, for the identification of the most promising first hits and also for a powerful database search for the identification of real epitopes and receptors, respectively.

For AFFI, it is expected that its reliability increases for a larger underlying dataset, which hopefully will be available in the near future. Additionally, the question of the practicability of AFFI for sequence alignment purposes is open, but a very interesting candidate for an attempt should be the CDR of an antibody.

The approach of covering sequence space with a reduced set of amino acids emulates the approach of the immune response, which has the challenge to fight against any possible antigen and therefore has to scan the complete sequence space. The recombination of V(D)J genes covers sequence space coarsely, which corresponds to the reduced set. Another parallel to be seen is in the fine-tuning step in the germinal center as this is related to the enrichment with the remaining amino acids.

Selection on BCR sequences

The findings presented here demonstrate that affinity maturation of antibodies is even more complex than previously thought. It involves a whole spectrum of relevant mutations, not all of which relate to promoting antigen binding through affinity enhancement, but presumably to additional effects such as efficient protein folding and resistance to proteolysis. Most strikingly, it suggests that selection might also act on silent mutations and therefore apparently at the level of antibody mRNA. The findings imply a scenario of affinity maturation with an initial phase involving non-affinity related increase in B cell fitness, e.g., through mediating sufficient and stable BCR expression levels. This might be the prerequisite for a subsequent phase involving affinity increase of BCRs towards the antigen.

An experimental validation for the positive selection of silent mutations is still pending and would dramatically increase the impact of that tremendous result.

An ancillary result that was indispensable for the investigation of the selection process in germinal centers, was the new development of a reliable method for detection of clonal independency of BCR sequences. The even

more complicated task of detecting clonal *relatedness* remains to be satisfactorily solved, although there are many groups working on this problem. Another goal for the future is the establishment of a method that is able to find correlations between favored substitutions to discover evolutionary pathways.

This study paved the way for further investigations...

Bibliography

[1] H. Akashi and A. Eyre-Walker. Translational selection and molecular evolution. *Curr Opin Genet Dev*, 8(6):688–93, 1998.

[2] C. D. Allen, T. Okada, and J. G. Cyster. Germinal-center organization and cellular dynamics. *Immunity*, 27(2):190–202, 2007.

[3] D. Allen, T. Simon, F. Sablitzky, K. Rajewsky, and A. Cumano. Antibody engineering for the analysis of affinity maturation of an anti-hapten response. *Embo J*, 7(7):1995–2001, 1988.

[4] Ian Anderson. A first course in combinatorial mathematics. *Oxford University Press*, 1974.

[5] Heiko Andresen, Carsten Grötzinger, Kim Zarse, Oliver J Kreuzer, Eva Ehrentreich-Förster, and Frank F Bier. Functional peptide microarrays for specific and sensitive antibody diagnostics. *Proteomics*, 6(5):1376–1384, Mar 2006.

[6] D. J. Barlow, M. S. Edwards, and J. M. Thornton. Continuous and discontinuous protein antigenic determinants. *Nature*, 322(6081):747–748, 1986.

[7] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*, 57(1):289–300, 1995.

[8] C. Berek and C. Milstein. Mutation drift and repertoire shift in the maturation of the immune response. *Immunol Rev*, 96:23–41, 1987.

[9] C. Berek, A. Berger, and M. Apel. Maturation of the immune response in germinal centers. *Cell*, 67(6):1121–9, 1991.

[10] A. G. Betz, C. Rada, R. Pannell, C. Milstein, and M. S. Neuberger. Passenger transgenes reveal intrinsic specificity of the antibody hypermutation mechanism: clustering, polarity, and specific hot spots. *Proc Natl Acad Sci U S A*, 90(6):2385–8, 1993.

[11] P. Boisguerin, R. Leben, B. Ay, G. Radziwill, K. Moelling, L. Dong, and R. Volkmer-Engert. An improved method for the synthesis of cellulose membrane-bound peptides with free c termini is useful for pdz domain binding studies. *Chem Biol*, 11(4):449–59, 2004.

[12] B. Bose and S. Sinha. Problems in using statistical analysis of replacement and silent mutations in antibody genes for determining antigen-driven affinity selection. *Immunology*, 116(2):172–83, 2005.

[13] Andrew P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30: 1145–1159, 1997.

[14] David B Carlini and Wolfgang Stephan. In vivo introduction of unpreferred synonymous codons into the drosophila adh gene results in reduced levels of adh protein. *Genetics*, 163(1):239–243, Jan 2003.

[15] Bruno Cernuschi-Frías. A combinatorial generalization of the stirling numbers of the second kind. *Proceedings of the 8th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, 2:593–596, 2001.

[16] J. V. Chamary and L. D. Hurst. Evidence for selection on synonymous mutations affecting stability of mrna secondary structure in mammals. *Genome Biol*, 6(9):R75, 2005.

[17] H. S. Chan. Folding alphabets. *Nature Structural Biology*, 6(11):994–996, 1999.

[18] H. S. Chan and K. A. Dill. Intrachain loops in polymers - effects of excluded volume. *Journal of Chemical Physics*, 90(1):492–509, 1989.

[19] Z. Chen, S. B. Koralov, M. Gendelman, M. C. Carroll, and G. Kelsoe. Humoral immune responses in cr2-/- mice: enhanced affinity maturation but impaired antibody persistence. *J Immunol*, 164(9):4522–32, 2000.

[20] L. T. Chong, Y. Duan, L. Wang, I. Massova, and P. A. Kollman. Molecular dynamics and free-energy calculations applied to affinity maturation in antibody 48g7. *Proc Natl Acad Sci U S A*, 96(25):14330–5, 1999.

[21] M. Cieplak, N. S. Holter, A. Maritan, and J. R. Banavar. Amino acid classes and the protein folding problem. *Journal of Chemical Physics*, 114(3):1420–1423, 2001.

[22] Louis A Clark, Skanth Ganesan, Sarah Papp, and Herman W T van Vlijmen. Trends in antibody sequence changes during the somatic hypermutation process. *J Immunol*, 177(1):333–340, Jul 2006.

[23] G. Cochrane, R. Akhtar, P. Aldebert, N. Althorpe, A. Baldwin, K. Bates, S. Bhattacharyya, J. Bonfield, L. Bower, P. Browne, M. Castro, T. Cox, F. Demiralp, R. Eberhardt, N. Faruque, G. Hoad, M. Jang, T. Kulikova, A. Labarga, R. Leinonen, S. Leonard, Q. Lin, R. Lopez, D. Lorenc, H. McWilliam, G. Mukherjee, F. Nardone, S. Plaister, S. Robinson, S. Sobhany, R. Vaughan, D. Wu, W. Zhu, R. Apweiler, T. Hubbard, and E. Birney. Priorities for nucleotide trace, sequence and annotation data capture at the ensembl trace archive and the embl nucleotide sequence database. *Nucleic Acids Res*, 36(Database issue): D5–12, 2008.

[24] A. Cumano and K. Rajewsky. Clonal recruitment and somatic mutation in the generation of immunological memory to the hapten np. *Embo J*, 5(10):2459–68, 1986.

[25] W. Dall'Acqua, E. R. Goldman, W. Lin, C. Teng, D. Tsuchiya, H. Li, X. Ysern, B. C. Braden, Y. Li, S. J. Smith-Gill, and R. A. Mariuzza. A mutational analysis of binding interactions in an antigen-antibody protein-protein complex. *Biochemistry*, 37(22):7981–91, 1998.

[26] A. R. Davidson, K. J. Lumb, and R. T. Sauer. Cooperatively folded proteins in random sequence libraries. *Nature Structural Biology*, 2 (10):856–864, 1995.

[27] M. Dayhoff, R. Schwartz, and B. Orcutt. A model of evolutionary change in protein. *Atlas Protein Seq. Struct.*, 5:345–352, 1978.

[28] W.L. DeLano. The pymol molecular graphics system. *http://www.pymol.org*, 2002.

[29] J. M. Di Noia and M. S. Neuberger. Molecular mechanisms of antibody somatic hypermutation. *Annu Rev Biochem*, 76:1–22, 2007.

[30] Liying Dong. Feinkartierung und charakterisierung linearer epitope von murinen monoklonalen antikörpern. *Ph.D. thesis, Charité Berlin*, 1998.

[31] J. Duan and M. A. Antezana. Mammalian mutation pressure, synonymous codon choice, and mrna degradation. *J Mol Evol*, 57(6):694–701, 2003.

[32] D. K. Dunn-Walters and J. Spencer. Strong intrinsic biases towards mutation and conservation of bases in human igvh genes during somatic hypermutation prevent statistical analysis of antigen selection. *Immunology*, 95(3):339–45, 1998.

[33] D. K. Dunn-Walters, A. Dogan, L. Boursier, C. M. MacDonald, and J. Spencer. Base-specific sequences that bias somatic hypermutation deduced by analysis of out-of-frame human igvh genes. *J Immunol*, 160 (5):2360–4, 1998.

[34] M. B. Eisen and P. O. Brown. Dna arrays for analysis of gene expression. *Methods Enzymol*, 303:179–205, 1999.

[35] Ferenc Evanics, Irina Bezsonova, Joseph Marsh, Julianne L Kitevski, Julie D Forman-Kay, and R. Scott Prosser. Tryptophan solvent exposure in folded and unfolded states of an sh3 domain by 19f and 1h nmr. *Biochemistry*, 45(47):14120–14128, Nov 2006.

[36] E. Fehrenbach, D. Zieker, A. M. Niess, E. Moeller, S. Russwurm, and H. Northoff. Microarray technology–the future analyses tool in exercise physiology? *Exerc Immunol Rev*, 9:58–69, 2003.

[37] Frederic A Fellouse, Christian Wiesmann, and Sachdev S Sidhu. Synthetic antibodies from a four-amino-acid code: a dominant role for tyrosine in antigen recognition. *Proc Natl Acad Sci U S A*, 101(34): 12467–12472, Aug 2004.

[38] Frederic A Fellouse, Bing Li, Deanne M Compaan, Andrew A Peden, Sarah G Hymowitz, and Sachdev S Sidhu. Molecular recognition by a binary code. *J Mol Biol*, 348(5):1153–1162, May 2005.

[39] Frederic A Fellouse, Pierre A Barthelemy, Robert F Kelley, and Sachdev S Sidhu. Tyrosine plays a dominant functional role in the

paratope of a synthetic antibody derived from a four amino acid code. *J Mol Biol*, 357(1):100–114, Mar 2006.

[40] W. Flemming. Studien über Regeneration der Gewebe. *Arch Mikros Anat*, 24:355–, 1885.

[41] R. Frank. The spot-synthesis technique. synthetic peptide arrays on membrane supports–principles and applications. *J Immunol Methods*, 267(1):13–26, 2002.

[42] A. Furukawa, K. Furukawa, and T. Azuma. A landscape for the dynamics of an immune response. *Biochem Biophys Res Commun*, 319 (2):469–78, 2004.

[43] K. Furukawa, A. Akasako-Furukawa, H. Shirai, H. Nakamura, and T. Azuma. Junctional amino acids determine the maturation pathway of an antibody. *Immunity*, 11(3):329–38, 1999.

[44] K. Furukawa, H. Shirai, T. Azuma, and H. Nakamura. A role of the third complementarity-determining region in the affinity maturation of an antibody. *J Biol Chem*, 276(29):27622–8, 2001.

[45] J. P. Gallivan and D. A. Dougherty. Cation-pi interactions in structural biology. *Proc Natl Acad Sci U S A*, 96(17):9459–9464, Aug 1999.

[46] H. M. Geysen, R. H. Meloen, and S. J. Barteling. Use of peptide-synthesis to probe viral-antigens for epitopes to a resolution of a single amino-acid. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences*, 81(13):3998–4002, 1984.

[47] V. Giudicelli, D. Chaume, and M. P. Lefranc. Imgt/v-quest, an integrated software program for immunoglobulin and t cell receptor v-j and v-d-j rearrangement analysis. *Nucleic Acids Res*, 32(Web Server issue):W435–40, 2004.

[48] A. Goede, I. S. Jaeger, and R. Preissner. Superficial - surface mapping of proteins via structure-based peptide library design. *Bmc Bioinformatics*, 6:–, 2005.

[49] A. Gonzalez-Fernandez, S. K. Gupta, R. Pannell, M. S. Neuberger, and C. Milstein. Somatic mutation of immunoglobulin lambda chains: a segment of the major intron hypermutates as much as the complementarity-determining regions. *Proc Natl Acad Sci U S A*, 91 (26):12614–8, 1994.

[50] G. M. Griffiths, C. Berek, M. Kaartinen, and C. Milstein. Somatic mutation and the maturation of immune response to 2-phenyl oxazolone. *Nature*, 312(5991):271–5, 1984.

[51] R. Grunow, R. Giese, T. Porstmann, H. Doepel, K. Haensel, and R. Vonbaehr. Development and biological testing of human and murine monoclonal-antibodies against hiv antigens. *Zeitschrift Fur Klinische Medizin-Zkm*, 45(4):367–369, 1990.

[52] S. Henikoff and J. G. Henikoff. Amino-acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22):10915–10919, 1992.

[53] U. Hershberg and M. J. Shlomchik. Differences in potential for amino acid change after mutation reveals distinct strategies for kappa and lambda light-chain variation. *Proc Natl Acad Sci U S A*, 103(43):15963–8, 2006.

[54] Uri Hershberg, Mohamed Uduman, Mark J Shlomchik, and Steven H Kleinstein. Improved methods for detecting selection by mutation analysis of ig v region sequences. *Int Immunol*, 20(5):683–694, May 2008.

[55] U. Hoffmuller, T. Knaute, M. Hahn, W. Hohne, J. Schneider-Mergener, and A. Kramer. Evolutionary transition pathways for changing peptide ligand specificity and structure. *Embo J*, 19(18):4866–74, 2000.

[56] W. E. Hohne, G. Kuttner, S. Kiessig, G. Hausdorf, R. Grunow, K. Winkler, H. Wessner, E. Giessmann, R. Stigler, J. Schneider-Mergener, and et al. Structural base of the interaction of a monoclonal antibody against p24 of hiv-1 with its peptide epitope. *Mol Immunol*, 30(13): 1213–21, 1993.

[57] A. Inamine, Y. Takahashi, N. Baba, K. Miyake, T. Tokuhisa, T. Takemori, and R. Abe. Two waves of memory b-cell generation in the primary immune response. *Int Immunol*, 17(5):581–9, 2005.

[58] J. Jacob and G. Kelsoe. In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. ii. a common clonal origin for periarteriolar lymphoid sheath-associated foci and germinal centers. *J Exp Med*, 176(3):679–87, 1992.

[59] J. Jacob, R. Kassir, and G. Kelsoe. In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. i. the architecture and dynamics of responding cell populations. *J Exp Med*, 173(5): 1165–75, 1991.

[60] J. Jacob, J. Przylepa, C. Miller, and G. Kelsoe. In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. iii. the kinetics of v region mutation and selection in germinal center b cells. *J Exp Med*, 178(4):1293–307, 1993.

[61] E. A. Kabat, T. T. Wu, H. M. Perry, K. S. Gottesman, and C. Foeller. *Sequences of Proteins of Immunological Interest*. U. S. Government Printing Office, Bethesda, MD, 1991.

[62] S. Kanaya, Y. Yamada, M. Kinouchi, Y. Kudo, and T. Ikemura. Codon usage and trna genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with cg-dinucleotide usage as assessed by multivariate analysis. *J Mol Evol*, 53(4-5):290–8, 2001.

[63] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa. Aaindex: amino acid index database, progress report 2008. *Nucleic Acids Res*, 36(Database issue):D202–5, 2008.

[64] H. Kimoto, H. Nagaoka, Y. Adachi, T. Mizuochi, T. Azuma, T. Yagi, T. Sata, S. Yonehara, Y. Tsunetsugu-Yokota, M. Taniguchi, and T. Takemori. Accumulation of somatic hypermutation and antigen-driven selection in rapidly cycling surface ig+ germinal center (gc) b cells which occupy gc at a high frequency during the primary anti-hapten response in mice. *Eur J Immunol*, 27(1):268–79, 1997.

[65] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217 (5129):624–626, Feb 1968.

[66] M. Kimura. The neutral theory of molecular evolution: a review of recent evidence. *Jpn J Genet*, 66(4):367–386, Aug 1991.

[67] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 1983.

[68] S. H. Kleinstein and J. P. Singh. Why are there so few key mutant clones? the influence of stochastic selection and blocking on affinity maturation in the germinal center. *Int Immunol*, 15(7):871–84, 2003.

[69] C. Kocks and K. Rajewsky. Stable expression and somatic hypermutation of antibody v regions in b-cell developmental pathways. *Annu Rev Immunol*, 7:537–59, 1989.

[70] A. Kouranov, L. Xie, J. de la Cruz, L. Chen, J. Westbrook, P. E. Bourne, and H. M. Berman. The rcsb pdb information portal for structural genomics. *Nucleic Acids Res*, 34(Database issue):D302–5, 2006.

[71] G. Kraal, I. L. Weissman, and E. C. Butcher. Germinal centre b cells: antigen specificity and changes in heavy chain class expression. *Nature*, 298(5872):377–379, Jul 1982.

[72] A. Kramer, T. Keitel, K. Winkler, W. Stocklein, W. Hohne, and J. Schneider-Mergener. Molecular basis for the binding promiscuity of an anti-p24 (hiv-1) monoclonal antibody. *Cell*, 91(6):799–809, 1997.

[73] A. Kramer, U. Reineke, L. Dong, B. Hoffmann, U. Hoffmuller, D. Winkler, R. Volkmer-Engert, and J. Schneider-Mergener. Spot synthesis: observations and optimizations. *J Pept Res*, 54(4):319–27, 1999.

[74] K. F. Lau and K. A. Dill. A lattice statistical-mechanics model of the conformational and sequence-spaces of proteins. *Macromolecules*, 22 (10):3986–3997, 1989.

[75] C. E H Lee, B. Gaëta, H. R. Malming, M. E. Bain, W. A. Sewell, and A. M. Collins. Reconsidering the human immunoglobulin heavy-chain locus: 1. an evaluation of the expressed human ighd gene repertoire. *Immunogenetics*, 57(12):917–925, Jan 2006.

[76] M. P. Lefranc. Imgt, the international immunogenetics information system: a standardized approach for immunogenetics and immunoinformatics. *Immunome Res*, 1:3, 2005.

[77] M. P. Lefranc, C. Pommie, M. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, V. Thouvenin-Contet, and G. Lefranc. Imgt unique numbering for immunoglobulin and t cell receptor variable domains and ig superfamily v-like domains. *Dev Comp Immunol*, 27(1):55–77, 2003.

[78] T. P. Li, K. Fan, J. Wang, and W. Wang. Reduction of protein sequence complexity by residue grouping. *Protein Engineering*, 16(5):323–330, 2003.

[79] Y. Li, H. Li, F. Yang, S. J. Smith-Gill, and R. A. Mariuzza. X-ray snapshots of the maturation of an antibody response to a protein antigen. *Nat Struct Biol*, 10(6):482–8, 2003.

[80] X. Liu, D. Liu, J. Qi, and W. M. Zheng. Simplified amino acid alphabets based on deviation of conditional probability from random background. *Physical Review E*, 66(2):–, 2002.

[81] Y. J. Liu, G. D. Johnson, J. Gordon, and I. C. MacLennan. Germinal centres in t-cell-dependent antibody responses. *Immunol Today*, 13(1): 17–21, 1992.

[82] Y. J. Liu, O. de Bouteiller, and I. Fugier-Vivier. Mechanisms of selection and differentiation in germinal centers. *Curr Opin Immunol*, 9(2): 256–62, 1997.

[83] C. D. Livingstone and G. J. Barton. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput Appl Biosci*, 9(6):745–756, Dec 1993.

[84] I. S. Lossos, R. Tibshirani, B. Narasimhan, and R. Levy. The inference of antigen selection on ig genes. *J Immunol*, 165(9):5122–6, 2000.

[85] Y. F. Lu, M. Singh, and J. Cerny. Canonical germinal center b cells may not dominate the memory response to antigenic challenge. *Int Immunol*, 13(5):643–55, 2001.

[86] G. MacBeath and S. L. Schreiber. Printing proteins as microarrays for high-throughput function determination. *Science*, 289(5485):1760–1763, Sep 2000.

[87] I. C. MacLennan. Germinal centers. *Annu Rev Immunol*, 12:117–39, 1994.

[88] J. S. Mattick and I. V. Makunin. Non-coding rna. *Hum Mol Genet*, 15 Spec No 1:R17–29, 2006.

[89] M. G. McHeyzer-Williams, M. J. McLean, P. A. Lalor, and G. J. Nossal. Antigen-driven b cell differentiation in vivo. *J Exp Med*, 178(1):295–307, 1993.

[90] K. S. Midelfort, H. H. Hernandez, S. M. Lippow, B. Tidor, C. L. Drennan, and K. D. Wittrup. Substantial energetic improvement with minimal structural perturbation in a high affinity mutant antibody. *J Mol Biol*, 343(3):685–701, 2004.

[91] A. P. Minton. Effects of excluded surface area and adsorbate clustering on surface adsorption of proteins i. equilibrium models. *Biophys Chem*, 86(2-3):239–47, 2000.

[92] A. P. Minton. The influence of macromolecular crowding and macromolecular confinement on biochemical reactions in physiological media. *J Biol Chem*, 276(14):10577–80, 2001.

[93] A. P. Minton. Effects of excluded surface area and adsorbate clustering on surface adsorption of proteins. ii. kinetic models. *Biophys J*, 80(4): 1641–8, 2001.

[94] S. Miyazawa and R. L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology*, 256 (3):623–644, 1996.

[95] Mehdi Yousfi Monod, Véronique Giudicelli, Denys Chaume, and Marie-Paule Lefranc. Imgt/junctionanalysis: the first tool for the analysis of the immunoglobulin and t cell receptor complex v-j and v-d-j junctions. *Bioinformatics*, 20 Suppl 1:i379–i385, Aug 2004.

[96] L. R. Murphy, A. Wallqvist, and R. M. Levy. Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Engineering*, 13(3):149–152, 2000.

[97] P. Nieuwenhuis and D. Opstelten. Functional anatomy of germinal centers. *Am J Anat*, 170(3):421–435, Jul 1984.

[98] M. Oprea, L. G. Cowell, and T. B. Kepler. The targeting of somatic hypermutation closely resembles that of meiotic mutation. *J Immunol*, 166(2):892–899, Jan 2001.

[99] M. Or-Guil, N. Wittenbrink, A. A. Weiser, and J. Schuchhardt. Recirculation of germinal center b cells: a multilevel selection strategy for antibody maturation. *Immunol Rev*, 216:130–41, 2007.

[100] Jonathan U Peled, Fei Li Kuang, Maria D Iglesias-Ussel, Sergio Roa, Susan L Kalis, Myron F Goodman, and Matthew D Scharff. The biochemistry of somatic hypermutation. *Annu Rev Immunol*, 26:481–511, 2008.

[101] K. W. Plaxco, D. S. Riddle, V. Grantcharova, and D. Baker. Simplified proteins: minimalist solutions to the 'protein folding problem'. *Current Opinion in Structural Biology*, 8(1):80–85, 1998.

[102] R Development Core Team. R: A language and environment for statistical computing, 2008. URL `http://www.R-project.org`.

[103] M. D. Radmacher, G. Kelsoe, and T. B. Kepler. Predicted and inferred waiting times for key mutations in the germinal centre reaction: evidence for stochasticity in selection. *Immunol Cell Biol*, 76(4):373–81, 1998.

[104] K. Rajewsky. Clonal selection and learning in the antibody system. *Nature*, 381(6585):751–8, 1996.

[105] K. Rajewsky, I. Förster, and A. Cumano. Evolutionary and somatic selection of the antibody repertoire in the mouse. *Science*, 238(4830): 1088–1094, Nov 1987.

[106] L. Regan and W. F. Degrado. Characterization of a helical protein designed from 1st principles. *Science*, 241(4868):976–978, 1988.

[107] C. Reidys, P. F. Stadler, and P. Schuster. Generic properties of combinatory maps: neutral networks of rna secondary structures. *Bull Math Biol*, 59(2):339–397, Mar 1997.

[108] U. Reineke, R. Volkmer-Engert, and J. Schneider-Mergener. Applications of peptide arrays prepared by the spot-technology. *Curr Opin Biotechnol*, 12(1):59–64, 2001.

[109] U. Reineke, C. Ivascu, M. Schlief, C. Landgraf, S. Gericke, G. Zahn, H. Herzel, R. Volkmer-Engert, and J. Schneider-Mergener. Identification of distinct antibody epitopes and mimotopes from a peptide array of 5520 randomly generated sequences. *J Immunol Methods*, 267(1): 37–51, 2002.

[110] A. M. Resch, L. Carmel, L. Marino-Ramirez, A. Y. Ogurtsov, S. A. Shabalina, I. B. Rogozin, and E. V. Koonin. Widespread positive selection in synonymous sites of mammalian genes. *Mol Biol Evol*, 24(8): 1821–31, 2007.

[111] Ida Retter, Hans Helmar Althaus, Richard Münch, and Werner Müller. Vbase2, an integrative v gene database. *Nucleic Acids Res*, 33 (Database issue):D671–D674, Jan 2005.

[112] D. S. Riddle, J. V. Santiago, S. T. BrayHall, N. Doshi, V. P. Grantcharova, Q. Yi, and D. Baker. Functional rapidly folding proteins from simplified amino acid sequences. *Nature Structural Biology*, 4(10):805–809, 1997.

[113] William H Robinson, Carla DiGennaro, Wolfgang Hueber, Brian B Haab, Makoto Kamachi, Erik J Dean, Sylvie Fournel, Derek Fong, Mark C Genovese, Henry E Neuman de Vegvar, Karl Skriner, David L Hirschberg, Robert I Morris, Sylviane Muller, Ger J Pruijn, Walther J van Venrooij, Josef S Smolen, Patrick O Brown, Lawrence Steinman, and Paul J Utz. Autoantigen microarrays for multiplex characterization of autoantibody responses. *Nat Med*, 8(3):295–301, Mar 2002.

[114] I. B. Rogozin and M. Diaz. Cutting edge: Dgyw/wrch is a better predictor of mutability at g:c bases in ig hypermutation than the widely accepted rgyw/wrcy motif and probably reflects a two-step activation-induced cytidine deaminase-triggered process. *J Immunol*, 172(6): 3382–4, 2004.

[115] Ravit Saada, Moran Weinberger, Gitit Shahaf, and Ramit Mehr. Models for antigen receptor gene rearrangement: Cdr3 length. *Immunol Cell Biol*, 85(4):323–332, Jun 2007.

[116] Johannes Schuchhardt, Liying Dong, Ulrich Hoffmüller, Achim Kramer, Jens Schneider-Mergener, and Hanspeter Herzel. Peptide binding landscapes. In *German Conference on Bioinformatics*, pages 183–188, 2000.

[117] S. A. Shabalina, A. Y. Ogurtsov, and N. A. Spiridonov. A periodic pattern of mrna secondary structure created by the genetic code. *Nucleic Acids Res*, 34(8):2428–37, 2006.

[118] G. S. Shapiro, K. Aviszus, D. Ikle, and L. J. Wysocki. Predicting regional mutability in antibody v genes based solely on di- and trinucleotide sequence composition. *J Immunol*, 163(1):259–68, 1999.

[119] G. S. Shapiro, K. Aviszus, J. Murphy, and L. J. Wysocki. Evolution of ig dna sequence to target specific base positions within codons for somatic hypermutation. *J Immunol*, 168(5):2302–6, 2002.

[120] G. S. Shapiro, M. C. Ellison, and L. J. Wysocki. Sequence-specific targeting of two bases on both dna strands by the somatic hypermutation mechanism. *Mol Immunol*, 40(5):287–95, 2003.

[121] M. Shimoda, T. Nakamura, Y. Takahashi, H. Asanuma, S. Tamura, T. Kurata, T. Mizuochi, N. Azuma, C. Kanno, and T. Takemori. Isotype-specific selection of high affinity memory b cells in nasal-associated lymphoid tissue. *J Exp Med*, 194(11):1597–607, 2001.

[122] M. J. Shlomchik, A. H. Aucoin, D. S. Pisetsky, and M. G. Weigert. Structure and function of anti-dna autoantibodies derived from a single autoimmune mouse. *Proc Natl Acad Sci U S A*, 84(24):9150–4, 1987.

[123] M. J. Shlomchik, A. Marshak-Rothstein, C. B. Wolfowicz, T. L. Rothstein, and M. G. Weigert. The role of clonal selection and somatic mutation in autoimmunity. *Nature*, 328(6133):805–11, 1987.

[124] M. J. Shlomchik, S. Litwin, and M. G. Weigert. The influence of somatic mutation on clonal expansion. *Prog. Immunol., Proc. 7th Int. Cong. Immunol.*, 7:415–423, 1990.

[125] M. J. Shlomchik, P. Watts, M. G. Weigert, and S. Litwin. Clone: a monte-carlo computer simulation of b cell clonal expansion, somatic mutation, and antigen-driven selection. *Curr Top Microbiol Immunol*, 229:173–97, 1998.

[126] N. Sinha and R. Nussinov. Point mutations and sequence variability in proteins. redistributions of preexisting populations. *Proceedings of the National Academy of Sciences of the United States of America*, 98(6): 3139–3144, 2001.

[127] D. S. Smith, G. Creadon, P. K. Jena, J. P. Portanova, B. L. Kotzin, and L. J. Wysocki. Di- and trinucleotide target preferences of somatic mutagenesis in normal and autoreactive b cells. *J Immunol*, 156(7): 2642–52, 1996.

[128] A. D. Solis and S. Rackovsky. Optimized representations and maximal information in proteins. *Proteins-Structure Function and Genetics*, 38 (2):149–164, 2000.

[129] Y. Takahashi, P. R. Dutta, D. M. Cerasoli, and G. Kelsoe. In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. v. affinity maturation develops in two stages of clonal selection. *J Exp Med*, 187(6):885–95, 1998.

[130] Y. Takahashi, D. M. Cerasoli, J. M. Dal Porto, M. Shimoda, R. Freund, W. Fang, D. G. Telander, E. N. Malvey, D. L. Mueller, T. W. Behrens, and G. Kelsoe. Relaxed negative selection in germinal centers and impaired affinity maturation in bcl-xl transgenic mice. *J Exp Med*, 190 (3):399–410, 1999.

[131] Y. Takahashi, H. Ohta, and T. Takemori. Fas is required for clonal selection in germinal centers and the subsequent establishment of the memory b cell repertoire. *Immunity*, 14(2):181–92, 2001.

[132] Y. Takahashi, A. Inamine, S. Hashimoto, S. Haraguchi, E. Yoshioka, N. Kojima, R. Abe, and T. Takemori. Novel role of the ras cascade in memory b cell response. *Immunity*, 23(2):127–38, 2005.

[133] V. Tapia, J. Bongartz, M. Schutkowski, N. Bruni, A. Weiser, B. Ay, R. Volkmer, and M. Or-Guil. Affinity profiling using the peptide microarray technology: a case study. *Anal Biochem*, 363(1):108–18, 2007.

[134] J. Valjakka, A. Hemminki, S. Niemi, H. Soderlund, K. Takkinen, and J. Rouvinen. Crystal structure of an in vitro affinity- and specificity-matured anti-testosterone fab in complex with testosterone. improved affinity results from small structural changes within the variable domains. *J Biol Chem*, 277(46):44021–7, 2002.

[135] R. Volkmer-Engert, B. Ehrhard, J. Hellwig, A. Kramer, W. Höhne, and J. Schneider-Mergener. Preparation, analysis and antibody binding studies of simultaneously synthesized soluble and cellulose-bound hiv-1 p24 peptide epitope libraries. *LIPS*, 1:243–253, 1994.

[136] R. Volkmer-Engert, B. Hoffmann, and J. Schneider-Mergener. Stable attachment of the hmb-linker to continuous cellulose membranes for parallel solid phase spot synthesis. *Tetrahedron Lett.*, 38:1029–1032, 1997.

[137] S. D. Wagner and M. S. Neuberger. Somatic hypermutation of immunoglobulin genes. *Annu Rev Immunol*, 14:441–57, 1996.

[138] J. Wang and W. Wang. A computational approach to simplifying the protein folding alphabet. *Nature Structural Biology*, 6(11):1033–1038, 1999.

[139] Yan Wang, Katherine J L Jackson, William A Sewell, and Andrew M Collins. Many human immunoglobulin heavy-chain ighv gene polymorphisms have been reported in error. *Immunol Cell Biol*, 86(2):111–115, Feb 2008.

[140] G. J. Wedemayer, P. A. Patten, L. H. Wang, P. G. Schultz, and R. C. Stevens. Structural insights into the evolution of an antibody combining site. *Science*, 276(5319):1665–9, 1997.

[141] H. Wenschuh, H. Gausepohl, L. Germeroth, M. Ulbricht, H. Matuschewski, A. Kramer, R. Volkmer-Engert, N. Heine, T. Ast, D. Scharn, and J. Schneider-Mergener. Positionally addressable parallel synthesis on continuous membranes. *in: (H.Fenniri ed.) Combinatorial Chemistry: A Practical Approach, Oxford University Press, Oxford, UK*, pages 95–116, 2000.

[142] H. Wenschuh, R. Volkmer-Engert, M. Schmidt, M. Schulz, J. Schneider-Mergener, and U. Reineke. Coherent membrane supports for parallel microsynthesis and screening of bioactive peptides. *Biopolymers*, 55 (3):188–206, 2000.

[143] U. Wiedemann, P. Boisguerin, R. Leben, D. Leitner, G. Krause, K. Moelling, R. Volkmer-Engert, and H. Oschkinat. Quantification of pdz domain specificity, prediction of ligand affinity and rational design of super-binding peptides. *J Mol Biol*, 343(3):703–18, 2004.

[144] B. Zheng, Z. Z. Ozen, S. Cao, Y. Zhang, and S. Han. Cd4-deficient t helper cells are capable of supporting somatic hypermutation and affinity maturation of germinal center b cells. *Eur J Immunol*, 32(11): 3315–25, 2002.

Appendix A

AFFI tables

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	13																			
C	6	5																		
D	10	8	28																	
E	11	3	23	16																
F	7	1	8	8	20															
G	15	2	13	9	9	20														
H	6	0	10	6	11	5	8													
I	11	3	0	2	12	2	5	11												
K	7	3	1	3	4	5	13	6	18											
L	5	5	4	3	15	5	3	20	8	18										
M	4	2	5	5	7	2	4	5	7	7	3									
N	4	1	8	6	3	4	8	4	7	6	3	14								
P	17	4	6	7	7	7	9	8	7	7	8	8	25							
Q	5	2	11	10	4	4	5	5	8	2	7	7	10	10						
R	12	0	3	3	6	6	9	7	26	8	8	9	6	11	24					
S	16	6	10	8	3	11	3	5	4	5	4	4	10	7	8	12				
T	11	4	7	5	4	9	5	13	5	6	3	8	9	6	10	21	16			
V	13	3	0	4	6	4	5	21	7	13	4	4	5	4	8	6	16	16		
W	2	0	1	5	13	3	3	4	0	5	2	3	4	0	3	0	2	4	11	
Y	4	1	2	2	22	4	6	6	5	5	4	3	3	3	3	3	2	5	6	15

Table A.1: C^{15} contains the number of observed *conserving* substitutions for key residues with the flexibility 15 or less.

	A	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**
A	0																			
C	12	0																		
D	31	25	0																	
E	18	18	21	0																
F	26	24	40	28	0															
G	18	23	35	27	31	0														
H	15	13	26	18	17	23	0													
I	13	13	39	25	19	29	14	0												
K	24	20	45	31	34	33	13	23	0											
L	26	18	42	31	23	33	23	9	28	0										
M	12	6	26	14	16	21	7	9	14	14	0									
N	23	18	34	24	31	30	14	21	25	26	14	0								
P	21	26	47	34	38	38	24	28	36	36	20	31	0							
Q	18	13	27	16	26	26	13	16	20	26	6	17	25	0						
R	25	29	49	37	38	38	23	28	16	34	19	29	43	23	0					
S	9	11	30	20	29	21	17	18	26	25	11	22	27	15	28	0				
T	18	17	37	27	32	27	19	14	29	28	16	22	32	20	30	7	0			
V	16	18	44	28	30	32	19	6	27	21	15	26	36	22	32	22	16	0		
W	22	16	38	22	18	28	16	18	29	24	12	22	32	21	32	23	25	23	0	
Y	24	19	41	29	13	31	17	20	28	28	14	26	37	22	36	24	29	26	20	0

Table A.2: H^{15} contains the number of observed *harmful* substitutions for key residues with the flexibility 15 or less.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	1.00																			
C	0.33	1.00																		
D	0.24	0.24	1.00																	
E	0.38	0.14	0.52	1.00																
F	0.21	0.04	0.17	0.22	1.00															
G	0.45	0.08	0.27	0.25	0.23	1.00														
H	0.29	0.00	0.28	0.25	0.39	0.18	1.00													
I	0.46	0.19	0.01	0.07	0.39	0.06	0.26	1.00												
K	0.23	0.13	0.02	0.09	0.11	0.13	0.50	0.21	1.00											
L	0.16	0.22	0.09	0.09	0.39	0.13	0.12	0.69	0.22	1.00										
M	0.25	0.25	0.16	0.26	0.30	0.09	0.36	0.36	0.33	0.33	1.00									
N	0.15	0.05	0.19	0.20	0.09	0.12	0.36	0.16	0.22	0.19	0.18	1.00								
P	0.45	0.13	0.11	0.17	0.16	0.16	0.27	0.22	0.16	0.16	0.29	0.21	1.00							
Q	0.22	0.13	0.29	0.38	0.13	0.13	0.28	0.24	0.29	0.07	0.54	0.29	0.29	1.00						
R	0.32	0.00	0.06	0.08	0.14	0.14	0.28	0.20	0.62	0.19	0.30	0.24	0.12	0.32	1.00					
S	0.64	0.35	0.25	0.29	0.09	0.34	0.15	0.22	0.13	0.17	0.27	0.15	0.27	0.32	0.22	1.00				
T	0.38	0.19	0.16	0.16	0.11	0.25	0.21	0.48	0.15	0.18	0.16	0.27	0.22	0.23	0.25	0.75	1.00			
V	0.45	0.14	0.01	0.13	0.17	0.11	0.21	0.78	0.21	0.38	0.21	0.13	0.12	0.15	0.20	0.21	0.50	1.00		
W	0.08	0.00	0.03	0.19	0.42	0.10	0.16	0.18	0.01	0.17	0.14	0.12	0.11	0.01	0.09	0.01	0.07	0.15	1.00	
Y	0.14	0.05	0.05	0.06	0.63	0.11	0.26	0.23	0.15	0.15	0.22	0.10	0.08	0.12	0.08	0.11	0.06	0.16	0.23	1.00

Table A.3: AFFI[15] contains the probabilities of *conserving* substitutions. The estimation is based on all key residues with flexibility 15 or less.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0%																			
C	15%	0%																		
D	11%	13%	0%																	
E	10%	25%	7%	0%																
F	15%	42%	14%	11%	0%															
G	7%	35%	11%	14%	12%	0%														
H	16%	0%	9%	13%	9%	19%	0%													
I	9%	20%	0%	29%	10%	31%	21%	0%												
K	14%	21%	43%	23%	19%	15%	9%	14%	0%											
L	20%	16%	19%	22%	8%	19%	23%	5%	13%	0%										
M	21%	24%	14%	13%	13%	32%	16%	17%	11%	12%	0%									
N	18%	40%	11%	15%	20%	22%	12%	18%	12%	15%	20%	0%								
P	10%	19%	18%	14%	12%	15%	11%	14%	15%	13%	11%	13%	0%							
Q	17%	25%	10%	9%	17%	20%	13%	16%	11%	27%	10%	11%	10%	0%						
R	10%	0%	28%	22%	17%	16%	10%	15%	5%	13%	14%	11%	16%	10%	0%					
S	6%	14%	11%	12%	19%	10%	22%	16%	19%	14%	15%	18%	12%	14%	13%	0%				
T	11%	17%	13%	17%	17%	13%	19%	8%	16%	15%	23%	13%	13%	15%	12%	3%	0%			
V	7%	20%	0%	24%	14%	18%	13%	5%	13%	9%	15%	19%	21%	20%	12%	15%	6%	0%		
W	29%	0%	49%	15%	9%	24%	22%	19%	0%	14%	30%	25%	19%	0%	22%	0%	27%	18%	0%	
Y	20%	41%	25%	32%	5%	17%	13%	13%	14%	17%	14%	20%	27%	22%	18%	20%	25%	16%	13%	0%

Table A.4: Coefficient of variation for each matrix entry - determined via bootstrapping.

l	k	p_1																					
1	2	0.63	ACDEFGHIKLMNPQRSTVV	W																			
1	3	0.76	ADEFGHIKLMNPQRSTVY	C	W																		
1	4	0.82	ADEFGHIKLMNPQRSTV	C	W	Y																	
1	5	0.86	AEFGHIKLMNPQRSTV	C	D	W	Y																
1	6	0.88	AEFHIKLMNPQRSTV	C	D	G	W	Y															
1	7	0.90	AFHIKLMNPQRSTV	C	D	E	G	W	Y														
1	8	0.92	AFHIKLMPQRSTV	C	D	E	G	N	W	Y													
1	9	0.93	AHIKLMPQRSTV	C	D	E	F	G	N	W	Y												
1	10	0.94	AHIKLMQRSTV	C	D	E	F	G	N	P	W	Y											
1	11	0.94	AHIKMQRSTV	C	D	E	F	G	L	N	P	W	Y										
1	12	0.95	AILMPQSTV	C	D	E	F	G	H	K	N	R	W	Y									
1	13	0.96	AILMQSTV	C	D	E	F	G	H	K	N	P	R	W	Y								
1	14	0.96	AILMSTV	C	D	E	F	G	H	K	N	P	Q	R	W	Y							
1	15	0.97	AILSTV	C	D	E	F	G	H	K	M	N	P	Q	R	W	Y						
1	16	0.97	AISTV	C	D	E	F	G	H	K	L	M	N	P	Q	R	W	Y					
1	17	0.98	AITV	C	D	E	F	G	H	K	L	M	N	P	Q	R	S	W	Y				
1	18	0.99	A	C	D	E	F	G	H	IV	K	L	M	N	P	Q	R	ST	W	Y			
1	19	0.99	A	C	D	E	F	G	H	IV	K	L	M	N	P	Q	R	S	T	W	Y		
1	20	1.00	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	

Table A.5: Optimal amino acid partitions for AFFI[15] estimated via simulated annealing for group size ($l = 1$) and several numbers of groups (k) according to p_1, Equation (4.15) - proximity.

l	k	p_1										
2	2	0.57	ACDEFGHKLMNPQRSTWY	IV								
2	3	0.67	ACDEFGHKLMNPQRWY	IV	ST							
2	4	0.71	ACDEGHKLMNPQRW	FY	IV	ST						
2	5	0.73	ACDEGHLMNPQW	FY	IV	KR	ST					
2	6	0.74	ACDEGHLNPW	FY	IV	KR	MQ	ST				
2	7	0.74	ACGHLNPW	DE	FY	IV	KR	MQ	ST			
2	8	0.74	AG	CHLNPW	DE	FY	IV	KR	MQ	ST		
2	9	0.74	AG	CLPW	DE	FY	HN	IV	KR	MQ	ST	
2	10	0.75	AG	CL	DE	FY	HN	IV	KR	MQ	PW	ST
3	2	0.50	ACDEFGHKMNPQRSTWY	ILV								
3	3	0.57	AST	CDEFGHKMNPQRWY	ILV							
3	4	0.60	AST	CDEGHKMNPQR	FWY	ILV						
3	5	0.61	AST	CDEGMNPQ	FWY	HKR	ILV					
3	6	0.61	AST	CMNPQ	DEG	FWY	HKR	ILV				
4	2	0.44	ACDEFGHKMNPQRSWY	ILTV								
4	3	0.49	AGST	CDEHKMNPQRWY	FILV							
4	4	0.51	ACDEGNPS	FHWY	ILTV	KMQR						
4	5	0.53	ACPS	DEMQ	FGWY	HKNR	ILTV					
5	2	0.42	AISTV	CDEFGHKLMNPQRWY								
5	3	0.45	AISTV	CDEFGLNPWY	HKMQR							
5	4	0.47	AGSTV	CDEMQ	FILWY	HKNPR						
6	2	0.39	AILSTV	CDEFGHKMNPQRWY								
6	3	0.43	ACDEGPST	FILVWY	HKMNQR							
7	2	0.37	ACILSTV	DEFGHKMNPQRWY								
8	2	0.35	ACILPSTV	DEFGHKMNQRWY								
9	2	0.34	ACDEGPQST	FHIKLMNRVWY								
10	2	0.34	ACDEGNPQST	FHIKLMRVWY								

Table A.6: Optimal amino acid partitions for AFFI^{15} estimated via simulated annealing for several group sizes ($l > 1$) and numbers of groups (k) according to p_1, Equation (4.15) - proximity.

l	k	p_2																			
1	2	0.34	ACDEGNPQST	FHIKLMRVWY																	
1	3	0.42	ACDEGST	FILVWY	HKMNPQR																
1	4	0.48	AIPSTV	CDEG	FLWY	HKMNQR															
1	5	0.53	ACPS	DEMQ	FGWY	HKNR	ILTV														
1	6	0.59	AGST	CDE	FWY	HKR	ILV	MNPQ													
1	7	0.64	AST	CN	DEG	FWY	HKR	ILV	MPQ												
1	8	0.68	AST	CP	DEG	FWY	HN	ILV	KR	MQ											
1	9	0.72	AST	CP	DE	FY	GW	HN	ILV	KR	MQ										
1	10	0.76	AG	CP	DE	FY	HN	ILV	KR	MQ	ST	W									
1	11	0.79	AG	C	DE	FY	HN	ILV	KR	MQ	P	ST	W								
1	12	0.81	AG	C	DE	FY	HN	IV	KR	L	MQ	P	ST	W							
1	13	0.84	AG	C	DE	FY	H	IV	KR	L	MQ	N	P	ST	W						
1	14	0.86	A	C	DE	FY	G	H	IV	KR	L	MQ	N	P	ST	W					
1	15	0.89	A	C	D	E	FY	G	H	IV	KR	L	MQ	N	P	ST	W				
1	16	0.91	A	C	D	E	FY	G	H	IV	KR	L	M	N	P	Q	ST	W			
1	17	0.94	A	C	D	E	FY	G	H	IV	K	L	M	N	P	Q	R	ST	W		
1	18	0.96	A	C	D	E	F	G	H	IV	K	L	M	N	P	Q	R	ST	W	Y	
1	19	0.98	A	C	D	E	F	G	H	IV	K	L	M	N	P	Q	R	S	T	W	Y
1	20	1.00	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y

Table A.7: Optimal amino acid partitions for AFFI[15] estimated via simulated annealing for group size ($l = 1$) and several numbers of groups (k) according to p_2, Equation (4.16) - weighted proximity.

l	k	p_2										
2	2	0.34	ACDEGNPQST	FHIKLMRVWY								
2	3	0.42	ACDEGST	FILVWY	HKMNPQR							
2	4	0.48	AIPSTV	CDEG	FLWY	HKMNQR						
2	5	0.54	AGPST	CILV	DEMQ	FWY	HKNR					
2	6	0.59	AGST	CDE	FWY	HKNR	ILV	MPQ				
2	7	0.64	AST	CN	DEG	FWY	HKR	ILV	MPQ			
2	8	0.68	AST	CG	DE	FWY	HKR	ILV	MQ	NP		
2	9	0.72	AST	CP	DE	FY	GW	HN	ILV	KR	MQ	
2	10	0.75	AG	CL	DE	FY	HN	IV	KR	MQ	PW	ST
3	2	0.34	ACDEGNPQST	FHIKLMRVWY								
3	3	0.42	ACDEGST	FILVWY	HKMNPQR							
3	4	0.48	AIPSTV	CDEG	FLWY	HKMNQR						
3	5	0.54	AGPST	CILV	DEMQ	FWY	HKNR					
3	6	0.59	AGST	CDE	FWY	HKNR	ILV	MPQ				
4	2	0.34	ACDEGNPQST	FHIKLMRVWY								
4	3	0.42	ACDEGST	FILVWY	HKMNPQR							
4	4	0.48	AIPSTV	CDEG	FLWY	HKMNQR						
4	5	0.53	ACPS	DEMQ	FGWY	HKNR	ILTV					
5	2	0.34	ACDEGNPQST	FHIKLMRVWY								
5	3	0.42	ACDEGST	FILVWY	HKMNPQR							
5	4	0.47	AGSTV	CDEMQ	FILWY	HKNPR						
6	2	0.34	ACDEGNPQST	FHIKLMRVWY								
6	3	0.42	ACDEGST	FILVWY	HKMNPQR							
7	2	0.34	ACDEGNPQST	FHIKLMRVWY								
8	2	0.34	ACDEGNPQST	FHIKLMRVWY								
9	2	0.34	ACDEGNPQST	FHIKLMRVWY								
10	2	0.34	ACDEGNPQST	FHIKLMRVWY								

Table A.8: Optimal amino acid partitions for AFFI[15] estimated via simulated annealing for several group sizes ($l > 1$) and numbers of groups (k) according to p_2, Equation (4.16) - weighted proximity.

l	k	p_3																							
1	2	0.12	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y			
1	3	0.09	A	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y				
1	4	0.10	A	C	E	F	G	H	I	L	M	N	P	Q	R	S	T	V	Y						
1	5	0.11	A	E	F	G	H	I	K	L	M	N	P	Q	S	T	V	Y							
1	6	0.11	A	E	F	G	H	I	K	L	M	N	P	Q	S	T	V	W							
1	7	0.12	A	E	F	G	H	I	K	L	M	N	Q	S	T	V									
1	8	0.13	A	E	F	H	I	K	L	M	N	Q	S	T	V										
1	9	0.13	A	E	F	H	I	K	L	M	N	Q	S	T											
1	10	0.14	A	E	F	H	I	K	L	M	Q	S	T												
1	11	0.15	A	E	F	H	I	L	M	Q	S	T													
1	12	0.15	A	F	H	I	L	M	Q	S	T														
1	13	0.16	A	F	H	I	M	Q	S	T															
1	14	0.17	A	F	H	I	M	S	T																
1	15	0.18	A	F	H	M	S																		
1	16	0.18	A	H	I	M	S																		
1	17	0.19	A	H	I	M																			
1	18	0.20	A	I	M																				
1	19	0.20	A	I																					
1	20	0.21	A																						

Table A.9: Optimal amino acid partitions for AFFI[15] estimated via simulated annealing for group size ($l = 1$) and several numbers of groups (k) according to p_3, Equation (4.17) - distance.

l	k	p_3										
2	2	0.13	ACDEFGHIKLMNPQRSTV	WY								
2	3	0.11	AEFGHIKLMNPQRSTV	CD	WY							
2	4	0.12	AEFGHILMNPQSTV	CD	KR	WY						
2	5	0.13	AEFHILMNQSTV	CP	DG	KR	WY					
2	6	0.15	AFHILMQSTV	CG	DE	KR	NP	WY				
2	7	0.16	AFHIMQST	CG	DE	KR	LV	NP	WY			
2	8	0.17	AHIMQS	CL	DE	FY	GW	KR	NP	TV		
2	9	0.19	AHIM	CP	DE	FY	GW	KR	LV	NQ	ST	
2	10	0.20	AG	CL	DE	FY	HN	IV	KR	MQ	PW	ST
3	2	0.14	ACDEGHIKLMNPQRSTV	FWY								
3	3	0.14	AEGHIKLMPQRSTV	CDN	FWY							
3	4	0.14	AGHILMPQSTV	CDE	FWY	KNR						
3	5	0.16	AHIMQSTV	CLP	DEG	FWY	KNR					
3	6	0.18	AMQST	CNP	DEG	FWY	HKR	ILV				
4	2	0.15	ACDEHIKLMNPQRSTV	FGWY								
4	3	0.15	AHIKMNPQRSTV	CDEG	FLWY							
4	4	0.16	AIMPQSTV	CDEG	FLWY	HKNR						
4	5	0.18	ACPS	DEMQ	FGWY	HKNR	ILTV					
5	2	0.16	ACEHIKLMNPQRSTV	DFGWY								
5	3	0.16	AHKMPQRSTV	CDEGN	FILWY							
5	4	0.18	AGSTV	CDEMQ	FILWY	HKNPR						
6	2	0.16	ACHIKLMNPQRSTV	DEFGWY								
6	3	0.17	ACDEGST	FILVWY	HKMNPQR							
7	2	0.16	AHIKLMNPQRSTV	CDEFGWY								
8	2	0.17	ACDEGNPQST	FHIKLMRVWY								
9	2	0.17	ACDEGNPQST	FHIKLMRVWY								
10	2	0.17	ACDEGNPQST	FHIKLMRVWY								

Table A.10: Optimal amino acid partitions for AFFI^{15} estimated via simulated annealing for several group sizes ($l > 1$) and numbers of groups (k) according to p_3, Equation (4.17) - distance.

l	k	p_4																			
1	2	0.12	ACDEFGHIKLMNPQRSTVY	W																	
1	3	0.13	ADEFGHIKLMNPQRSTVY	C	W																
1	4	0.14	ADEFGHIKLMNPQRSTV	C	W	Y															
1	5	0.15	AEFGHIKLMNPQRSTV	C	D	W	Y														
1	6	0.16	AEFHIKLMNPQRSTV	C	D	G	W	Y													
1	7	0.16	AFHIKLMNPQRSTV	C	D	E	G	W	Y												
1	8	0.17	AFHIKLMPQRSTV	C	D	E	G	N	W	Y											
1	9	0.17	AHIKLMPQRSTV	C	D	E	F	G	N	W	Y										
1	10	0.18	AHIKLMQRSTV	C	D	E	F	G	N	P	W	Y									
1	11	0.18	AHIKMQRSTV	C	D	E	F	G	L	N	P	W	Y								
1	12	0.19	AILMPQSTV	C	D	E	F	G	H	K	N	R	W	Y							
1	13	0.19	AILMSTV	C	D	E	FY	G	H	K	N	P	Q	R	W						
1	14	0.19	AILSTV	C	D	E	FY	G	H	K	M	N	P	Q	R	W					
1	15	0.20	AILSTV	C	D	E	F	G	H	K	M	N	P	Q	R	W	Y				
1	16	0.20	AISTV	C	D	E	F	G	H	K	L	M	N	P	Q	R	W	Y			
1	17	0.20	AITV	C	D	E	F	G	H	K	L	M	N	P	Q	R	S	W	Y		
1	18	0.21	A	C	D	E	F	G	H	ILV	K	M	N	P	Q	R	S	T	W	Y	
1	19	0.21	A	C	D	E	F	G	H	IV	K	L	M	N	P	Q	R	S	T	W	Y
1	20	0.21	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y

Table A.11: Optimal amino acid partitions for AFFI^{15} estimated via simulated annealing for group size ($l = 1$) and several numbers of groups (k) according to p_4, Equation (4.18) - weighted distance.

l	k	p_4										
2	2	0.13	ACDEFGHIKLMNPQRSTV	WY								
2	3	0.15	AEFGHIKLMNPQRSTV	CD	WY							
2	4	0.16	AHIKLMNPQRSTV	CG	DE	FWY						
2	5	0.17	AHIKLMPQRSTV	CG	DE	FY	NW					
2	6	0.18	AHIKMQRSTV	CL	DE	FY	GW	NP				
2	7	0.18	ACILSTV	DE	FWY	GP	HN	KR	MQ			
2	8	0.19	AILSTV	CP	DE	FY	GW	HN	KR	MQ		
2	9	0.19	AST	CP	DE	FY	GW	HN	ILV	KR	MQ	
2	10	0.20	AG	CL	DE	FY	HN	IV	KR	MQ	PW	ST
3	2	0.14	ACDEGHIKLMNPQRSTV	FWY								
3	3	0.15	AGHIKLMNPQRSTV	CDE	FWY							
3	4	0.17	AHIKLMQRSTV	CNP	DEG	FWY						
3	5	0.18	ACILSTV	DEG	FWY	HKNR	MPQ					
3	6	0.18	AGST	CDE	FWY	HKNR	ILV	MPQ				
4	2	0.15	ACDEHIKLMNPQRSTV	FGWY								
4	3	0.16	AHIKMNPQRSTV	CDEG	FLWY							
4	4	0.17	ACILPSTV	DEMQ	FGWY	HKNR						
4	5	0.18	ACPS	DEMQ	FGWY	HKNR	ILTV					
5	2	0.16	ACEHIKLMNPQRSTV	DFGWY								
5	3	0.17	ACDEGPST	FILVWY	HKMNQR							
5	4	0.18	AGSTV	CDEMQ	FILWY	HKNPR						
6	2	0.16	ACHIKLMNPQRSTV	DEFGWY								
6	3	0.17	ACDEGPST	FILVWY	HKMNQR							
7	2	0.16	AHIKLMNPQRSTV	CDEFGWY								
8	2	0.17	ACDEGNPQST	FHIKLMRVWY								
9	2	0.17	ACDEGNPQST	FHIKLMRVWY								
10	2	0.17	ACDEGNPQST	FHIKLMRVWY								

Table A.12: Optimal amino acid partitions for AFFI^{15} estimated via simulated annealing for several group sizes ($l > 1$) and numbers of groups (k) according to p_4, Equation (4.18) - weighted distance.

l	k	p_5																			
1	2	0.44	ACDEGIKNPQRSTV	FHLMWY																	
1	3	0.51	ACDEGPS	FILMTVWY	HKNQR																
1	4	0.57	ACDEGPS	FWY	HKNQR	ILMTV															
1	5	0.63	ACGST	DEMNPQ	FWY	HKR	ILV														
1	6	0.67	ACGST	DE	FWY	HKR	ILV	MNPQ													
1	7	0.70	ACGST	DE	FWY	HKR	ILV	MNQ	P												
1	8	0.74	ACST	DE	FWY	G	HKR	ILV	MNQ	P											
1	9	0.77	ACST	DE	FWY	G	HKR	ILV	MQ	N	P										
1	10	0.80	AST	C	DE	FWY	G	HKR	ILV	MQ	N	P									
1	11	0.83	AST	C	DE	FY	G	HKR	ILV	MQ	N	P	W								
1	12	0.86	AST	C	DE	FY	G	H	ILV	KR	MQ	N	P	W							
1	13	0.88	AST	C	D	E	FY	G	H	ILV	KR	MQ	N	P	W						
1	14	0.91	AST	C	D	E	FY	G	H	ILV	KR	M	N	P	Q	W					
1	15	0.92	AST	C	D	E	FY	G	H	ILV	K	M	N	P	Q	R	W				
1	16	0.94	AST	C	D	E	F	G	H	ILV	K	M	N	P	Q	R	W	Y			
1	17	0.96	A	C	D	E	F	G	H	ILV	K	M	N	P	Q	R	ST	W	Y		
1	18	0.98	A	C	D	E	F	G	H	IV	K	L	M	N	P	Q	R	ST	W	Y	
1	19	0.99	A	C	D	E	F	G	H	IV	K	L	M	N	P	Q	R	S	T	W	Y
1	20	1.00	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y

Table A.13: Optimal amino acid partitions for AFFI^{15} estimated via simulated annealing for group size ($l = 1$) and several numbers of groups (k) according to p_5, Equation (4.19) - representatives.

l	k	p_5											
2	2	0.44	ACDEGIKNPQRSTV	FHLMWY									
2	3	0.51	ACDEGPS	FILMTVWY	HKNQR								
2	4	0.57	ACDEGPS	FWY	HKNQR	ILMTV							
2	5	0.63	ACGST	DEMNPQ	FWY	HKR	ILV						
2	6	0.67	ACGST	DE	FWY	HKR	ILV	MNPQ					
2	7	0.70	AGP	CST	DE	FWY	HKR	ILV	MNQ				
2	8	0.73	AGP	CST	DE	FWY	HN	ILV	KR	MQ			
2	9	0.74	AG	CP	DE	FWY	HN	ILV	KR	MQ	ST		
2	10	0.75	AG	CL	DE	FY	HN	IV	KR	MQ	PW	ST	
3	2	0.44	ACDEGIKNPQRSTV	FHLMWY									
3	3	0.51	ACDEGPS	FILMTVWY	HKNQR								
3	4	0.57	ACDEGPS	FWY	HKNQR	ILMTV							
3	5	0.63	ACGST	DEMNPQ	FWY	HKR	ILV						
3	6	0.66	ACST	DEG	FWY	HKR	ILV	MNPQ					
4	2	0.44	ACDEGIKNPQRSTV	FHLMWY									
4	3	0.51	ACDEGPS	FILMTVWY	HKNQR								
4	4	0.57	ACDEGPS	FHWY	ILMTV	KNQR							
4	5	0.61	ACST	DEGQ	FMWY	HKNR	ILPV						
5	2	0.44	ACDEGIKNPQRSTV	FHLMWY									
5	3	0.51	ACDEGPS	FILMTVWY	HKNQR								
5	4	0.56	ACGPS	DEFWY	HKNQR	ILMTV							
6	2	0.44	ACDEGIKNPQRSTV	FHLMWY									
6	3	0.51	ACDEGPS	FILTVWY	HKMNQR								
7	2	0.44	ACDEGHKPRS	FILMNQTVWY									
8	2	0.44	ACDEGHKPRS	FILMNQTVWY									
9	2	0.44	ACDEGHKPRS	FILMNQTVWY									
10	2	0.44	ACDEGHKPRS	FILMNQTVWY									

Table A.14: Optimal amino acid partitions for AFFI[15] estimated via simulated annealing for several group sizes ($l > 1$) and numbers of groups (k) according to p_5, Equation (4.19) - representatives.

Appendix B

Abbreviations

Abbreviation	Explanation
β-Ala	β-alanine
AA	amino acid
Ag	antigen
AUC	area under curve (ROC)
BCR	B cell receptor
BG	background signal intensity
BLU	Boehringer light units
BM	bone marrow
C-Term	carboxy-end in peptides or proteins
CDR	complementarity determining region
CGG	chicken γ-globulin
CM	cellulose membrane
DMF	dimethylformamide
DMSO	dimethylsulfoxide
DNA	deoxyribonucleic acid
Fab	fragment antigen binding
Fc	fragment crystallizable
Fmoc	9-fluorenylmethyloxycarbonyl
FPR	false positive rate
FQ	membrane functionality quotient
FR	framework region
Fv	variable fragment
GC	germinal center
Ig	immunoglobulin
IgH	Ig heavy chain
HPLC	high performance liquid chromatography

Abbreviation	Explanation
K_{dis}	dissociation constant
LT	lymphatic tissue
mAb	monoclonal antibody
MALDI-MS	matrix-assisted laser desorption ionization mass spectrometry
mRNA	messenger RNA
N-Term	amino-end in peptides or proteins
NP	4-hydroxy-3-nitrophenylacetyl
p_{IL}	conserving probability of a substitution between the amino acids I and L
$p_{[ILV]}$	conserving probability within the cluster I, L, V
PDB	Protein Data Bank
PID	protein interaction domain
pK_{dis}	negative decadic logarithm of K_{dis}
POD	peroxidase
R	replacement mutations
R/S	ratio of replacement (R) and silent mutations (S)
RNA	ribonucleic acid
ROC	receiver operator characteristic
S	silent mutations
SA	simulated annealing
SCM	standard cellulose membrane
SHM	somatic hypermutation
SI	signal intensity
SNR	signal to noise ratio
SPR	surface plasmon resonance
T-TBS	Tween-Tris buffered saline
TFA	trifluoro acetic acid
TPR	true positive rate
tRNA	transfer RNA
V region	variable region of Ig
V(D)J	variable (diversity) joining genes
V_H	variable region of the heavy chain
VH	V gene sequence of the heavy chain
V_L	variable region of the light chain
VL	V gene sequence of the light chain

Amino acids were abbreviated in the common way. Depending on the context the long (three-letters code) or the short (one-letter code) abbreviation was

used.
Nucleic acids were abbreviated in the one-letter code (A, C, G, T, U).

Acknowledgments

I want to express my gratitude to all those who have given me valuable scientific and personal support during this work.

I want to thank **Cornelius Frömmel** for accommodating me in his research group and introducing me to the world of proteins and amino acids. He always had great ideas for solving scientific as well as bureaucratic problems. Unfortunately, we never found time for a nice beach volleyball match.

Michal Or-Guil who introduced me to the world of systems immunology, which was the perfect continuation for my work. She had already co-assisted me in the first phase of my work. I want to thank her for the many fruitful discussions, where lots of the ideas presented here were born.

Rudolf Volkmer for experimental support.

Johannes Schuchhardt was the man, who always found the time to discuss tools and methods in statistics with me.

Achim Kramer for kind and helpful discussions.

Nicole Wittenbrink for countless discussions about the biology of the immune system and furthermore, for the invaluable support for writing and preparing figures. I would also like to thank **Atijeh Valai** and **Andrej I. Schmelzer** for helpful literature research. **Lei Zhang** for intensive and good teamwork and for introducing me to the secrets of the chinese hotpot. And, of course, **Juliane "Bongie" Bongartz** for her help with so many things: chemistry questions, preparing figures, writing, proofreading, musical taste, laughing and being there. Thanks to the other fellow members of the Systems Immunology group, **Christin**, **Nicole B.**, **Henning** and **Tom** for their cooperation and patience.

A very special thank you is for **Astrid Leichsenring**, who executed nearly all the experiments presented in this work. **Ines Kretzschmar**, **Christiane Landgraf** and **Victor Tapia** also gave me great experimental support.

I want to thank the groups "Structural Bioinformatics Group" with its head **Robert Preissner**, "Molekulare Bibliotheken" with its head **Rudolf Volkmer** and the little fraction of the AG Geginat for discussion, the time spent together, the steady help with words and deeds and the always warm atmosphere. Many of the group members evolved into friends.

Melanie, **Carsten** and **Victor G.** for proofreading.
Ralf, **Jan**, **Stefan**, **Melanie**, **Elke** and **Ines** for perfect atmosphere in our homey "flat share". The same counts for **Bodo**, **Svenja** and **Anja** in another flat.
Rudi, **Astrid**, **Jana**, **Christiane**, **Ines**, **Prisca**, **Judith**, **Zerrin**, **Bernhard**, **Marc**, **Lars**, **Michi**, **Mathias** and **Jens** for the good spirit in your lab and for never giving up trying to beat me in tabletop soccer.

I want to deeply thank **Grit Kasper** for proofreading, helpful discussions and several dinners.

I thank my mother for her years-long support.

A big thank you is for my wife *in spe*, **Katrin**, for her patience and personal support during the preparation of this thesis.

Selbständigkeitserklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit mit dem Titel „Amino acid substitutions in protein binding: A study for peptides and antibodies" selbständig und ohne Hilfe Dritter angefertigt habe. Sämtliche Hilfsmittel, Hilfen sowie Literaturquellen sind als solche kenntlich gemacht.

Ich habe mich anderweitig nicht um einen Doktorgrad beworben und besitze auch keinen entsprechenden Doktorgrad.

Mir ist die dem Verfahren zugrunde liegende Promotionsordnung bekannt.

Berlin, den

Armin A. Weiser